Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF MATHEMATICS

MSCI RESEARCH PROJECT

A causal approach for control set identification in empirical finance studies

Author:
Quinn Wang

Supervisor(s):
Dr. Simone Cenci

Submitted in partial fulfillment of the requirements for the MSci in Mathematics at Imperial College London

June 24, 2024

Abstract

Regression models in empirical finance studies typically use control variables without clear justification, leading to biased results and reduced interpretability. In this study, we focus on an emerging literature documenting the relationships between emissions and stock returns. We propose a systematic approach using Structural Causal Modeling (SCM) to identify a sufficient control set that avoids these biases.

Plagiarism statement

The work contained in this thesis is my own work unless otherwise stated.

Signature: Quinn Wang
Date: June 24, 2024

Contents

Chapter 1

Introduction

The relationship between carbon emissions and stock returns is a widely discussed issue in recent empirical finance studies (In et al. 2017, Bolton & Kacperczyk 2021, Aswani et al. 2024). The central question addressed here is: *Do investors care about carbon emissions?* (which is also the title of Bolton & Kacperczyk (2021)). While Bolton & Kacperczyk (2021) finds a positive association between stock returns and emissions measured in total emission level and emission growth, a more recent study Aswani et al. (2024) argues that total emissions and emission growth rates are primarily indicators of firm sales and sales growth rate, rather than capturing the firm's actual pollution levels. Both studies employ regression analysis to investigate these relationships.

In this report, we propose to use a causal framework to answer a qualitative question and a quantitative question: (1) Does the association between emissions and returns truly exist? (2). What is the magnitude of the associations between emissions and returns?

We first perform the main regression done in Bolton & Kacperczyk (2021) as the baseline regression. This regression (Eq.3.1) takes stock returns as the dependent variable and regresses it on emissions and a set of control variables (which are firm characteristics). These control variables are known to predict stock returns, aiming to isolate the effect of emissions on stock returns, and are used without justification.

However, as shown in the examples for collider bias and mediator bias in section 2.3, including control variables based on an untested belief on what influences the dependent variable can lead to biased results. Thus, the second question to answer translates to how to select the right control variables that are sufficient and do not induce bias in regression results.

Structural causal modeling (SCM) provides a systematic approach to answer this question. We first formulate our hypothesis on the causal story between emissions, stock returns, and other firm characteristics. This hypothesis considering causal structures between variables can be represented on a causal graph (which is the graphical representation of the associated structural causal model). A causal graph, which is a directed acyclic graph (DAG), entails a set of conditional independencies that can be tested on empirical data (Pearl 1995). Thus, we can validate our causal model through conditional independence tests. We answer the first question in section 4.2.3 when conducting conditional independence tests between emissions and stock returns. The causal model I developed and validated is shown in Figure 4.5. Finally, by applying the backdoor criterion to the validated causal model, we obtain the control set we want in the second question. We then re-estimate the magnitude of associations between emission and stock returns. The re-estimated associations reveal results that partly differ from those of the baseline regression. These differences underscore the potential biases present in the baseline estimates.

Note that the SCM approach focuses on causal relationships instead of statistical correlations in standard regression models. Also, a structural causal model is, in theory, falsifiable (Cenci & Kealhofer 2022, Pearl et al. 2016). We can test our hypothesis by considering causal relationships in data and explicitly correct our causal model if there are misspecifications. In contrast, regression models are not falsifiable (Taagepera 2008). Note that we use regression as an estimation method instead of a statistical model in this report. All the code used in this manuscript can be found in Appendix B.

Chapter 2

Background

In this chapter, we will go through an overview of structural causal modeling. Structural causal modeling is a representation of the data-generating mechanism between variables (Pearl et al. 2016). Through the model, we tell a causal story between the variables, i.e. how the data were generated according to the causality among the variables behind the scenes. Each structural causal model is associated with a graphical model, which entails a set of conditional independencies. We can test those conditional independencies on data ex-ante and thus validate our structural causal model (a structural causal model is falsifiable (Cenci & Kealhofer 2022)). This is in contrast to the regression models used in many current empirical finance studies, which can only be interpreted and evaluated after the fact (ex-post).

As preliminary, some basic concepts in graph theory are introduced in Appendix A.1. Readers who are familiar with the topic can safely skip this part.

2.1 Structural Causal Models

In order to deal with the questions of causality more rigorously, we introduce structural causal models or SCM. The goal of a structural causal model is to represent causal relationships among a set of variables, i.e., a structural causal model is a mathematical formalization of a causal story behind a data set (Cenci & Kealhofer 2022). It describes how the causal nature assigns values to variables of interest.

Formally, a structural causal model consists of two sets of variables U and V, and a set of functions f that assigns each variable in V based on the values of the other variables in the model (Pearl et al. 2016). Here we build our first definition of causation:

Definition 1. *(Pearl et al. 2016) A variable X is a **direct cause** of a variable Y if X appears in the function that assigns Y's value. X is a **cause** of Y if it is a direct cause of Y, or of any cause of Y.*

In other words, a structural causal model is a set of structural equations with a well-defined direction of causation. It is important to note that inverting the equations in a structural causal model leads to a completely different, and possibly wrong, causal story (Cenci & Kealhofer 2022).

Every SCM is associated with a graphical causal model, which can be referred to informally as a "graphical model" or "causal graph" (Pearl et al. 2016). We primarily deal with SCM with graphical models which are directed acyclic graphs (DAGs)[1]. The nodes in a graphical model represent the variables in U and V, and a set of edges between the nodes represents the

[1] Cycles are not allowed in a causal graph.

functions in f. The function f_X for variable X takes the values of the causes of X as input and assigns the value of X. From the relationship between SCMs and graphical causal models, we give the second graphical definition of causation:

Definition 2. *(Pearl et al. 2016) In a graphical model, if a variable X is the child of another variable Y, then Y is a* ***direct cause*** *of X; if X is a descendant of Y, then Y is a* ***potential cause*** *of X.*

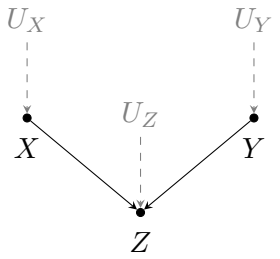


Figure 2.1 Graphical model of firm carbon emission level based on firm size and sales, with X indicating the firm size, Y indicating sales of the firm, and Z indicating the amount of carbon emission. The gray nodes U_X, U_Y, and U_Z are unobserved exogenous variables, and the gray edges are unobserved edges. This is a collider structure which we will introduce in section 2.2.

Consider the example in Figure 2.1, the graphical model in Figure 2.1 is the only graphical model associated with the following SCM:

$$U = \{U_X, U_Y, U_Z\}, \quad V = \{X, Y, Z\}, \quad f = \{f_X, f_Y, f_Z\} \tag{2.1a}$$

$$X = f_X(U_X), \quad Y = f_Y(U_Y) \quad Z = f_Z(X, Y, U_Z) \tag{2.1b}$$

We can observe from both the graphical model and the associated SCM that both the firm size (X) and sales of the firm (Y) are direct causes of the firm's carbon emission (Z). The variables in U are called *exogenous variables*, which means they are unobserved external causes to the model, and we choose, for whatever reason, not to explain how they are caused. Therefore, exogenous variables have no cause and are always root nodes in graphs. The variables in V are *endogenous variables*. Every endogenous variable is a descendant of at least one exogenous variable (Pearl et al. 2016). We differentiate the unobserved exogenous variables from others by plotting them as gray nodes and using dotted gray lines for unobserved edges connected to them. From section 2.4 and onwards, we will only show the exogenous variables which are the common causes of multiple endogenous variables in the graph. As we will see later, it is crucial to understand how some unobserved exogenous variables enter into the model.

While the structural equations quantitatively define the causal relations, a causal graph provides a more intuitive image of qualitative causalities entailed by the underlying variables. Furthermore, in most cases, our knowledge about causal relationships is not quantitative, but qualitative (Pearl et al. 2016), as elaborated in a graphical model. Therefore, from now on, we will always represent a SCM by its associated graphical model.

2.2 Graphical Models

There are three fundamental configurations in graphical causal models: *chains* (Figure 2.2), *forks* (Figure 2.3), and *colliders* (Figure 2.4). We can build any arbitrary complicated causal graph (which are DAGs) using these three building blocks.

Each of the three building blocks (and each causal graph or structural causal model in general) entails a particular set of conditional independencies. Hence, a structural causal model

is often falsifiable as each of these independencies can, in theory, be tested (and rejected) from empirical data (Pearl 1995, Cenci & Kealhofer 2022).

For example, a *chain* (Figure 2.2) entails $X \perp\!\!\!\perp Z \mid Y$, i.e. X and Z are independent conditional on Y . When we fix the value of $Y = a$, different values of X will not affect the value of Y, and the value of Z depends only on U_Z and $Y = a$. Hence a change in X does not change the value of Z, i.e. X and Z are independent, conditional on Y. Note that we assume the unobserved variables U_X, U_Y, and U_Z are independent of each other. If, for example, U_X were a cause of U_Y, then variations in X can be associated with variations in Y through the exogenous variables U_X and U_Y, hence X and Z are not necessarily conditionally independent given Y.

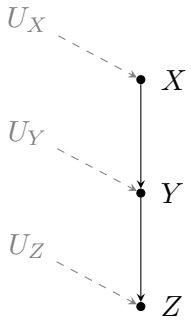


Figure 2.2 Chain

A *fork* in Figure 2.3 contains a common cause Y of X and Z. The conditional independencies entailed in a fork structure are the same as the chain structure, i.e. $X \perp\!\!\!\perp Z \mid Y$.

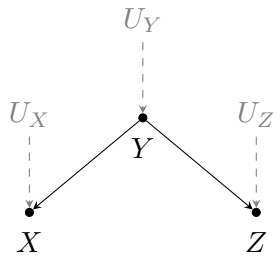


Figure 2.3 Fork

In Figure 2.4, a *collider* contains Y as the common effect of two direct causes X and Z, where Y is called the collision node. Assuming independence of U_X, U_Y, and U_Z, we have $X$and Z are unconditionally independent, i.e. $X \perp\!\!\!\perp Z$.

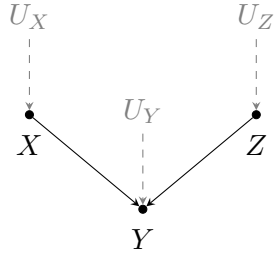


Figure 2.4 Collider

Therefore, if we believe that the causal relationships among variables follow a collider struc-

ture (like the one in Figure 2.4), we can test our hypothesis on data by examining the condition $X \perp\!\!\!\perp Z \mid Y$. If, however, we find that $X \perp\!\!\!\perp Z \mid Y$ holds true instead of $X \perp\!\!\!\perp Z$, we should reject our initial hypothesis. This result would suggest that the empirical data are more consistent with a chain or fork structure (Figure 2.2), indicating that our initial model does not accurately represent the underlying causal relationships.

Conditional independence testing on the empirical data helps us validate and reconstruct our SCM models. However, this approach has two important limitations. Firstly, some conditional independence conditions cannot be tested on data (Cenci & Kealhofer 2022). For instance, we cannot test the conditional independence conditions involving an unobserved variable. Secondly, while each SCM is associated with a unique causal graph, multiple causal graphs can entail the same set of conditional independencies. For example, a chain in Figure 2.2 and a fork in Figure 2.3 or two chains with the reverse direction of causation. In other words, exploiting the conditional independence conditions in a graph can only give us solutions up to a Markov equivalence class (Zhang et al. 2012). More details will be discussed in section 2.4.

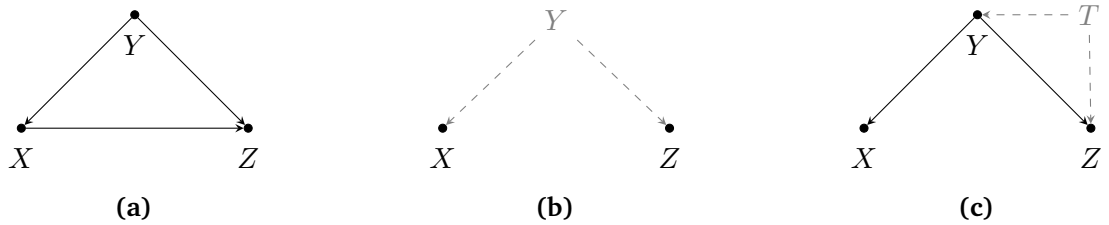


Figure 2.5 Examples of causal graphs with identifiability problems (Cenci & Kealhofer 2022). The independent exogenous variables that only influence one endogenous variable are hidden from the graph for simplicity. Panel (a) illustrates a graph that does not imply any conditional independence. Panel (b) illustrates a graph where we cannot measure the conditional independence condition $X \perp\!\!\!\perp Z \mid Y$ because the common cause Y is unobserved. Panel (c) implies more conditional independencies that we can test because of the unobserved variable T.

In real-world applications, the SCMs and their associated causal graphs can be much more complex than the three fundamental structures introduced above. Consequently, identifying conditional independencies directly from the graphs is often not straightforward. Fortunately, there exists a robust methodology for determining whether variables in a graph are conditionally (or unconditionally) dependent or independent. *d-separation* (where d stands for "directional") is a criterion that can be applied to a graphical model of any complexity to predict conditional independencies that are shared by all data sets generated by that graph (Pearl et al. 2016).

Definition 3. ***(d-separation)*** *(Pearl et al. 2016) A path p is blocked by a set of nodes Z if and only if*

1. *p contains a chain of nodes $A \to B \to C$ or a fork $A \leftarrow B \to C$ such that the middle node B is in Z (i.e., B is conditioned on), or*
2. *p contains a collider $A \to B \leftarrow C$ such that the collision node B is not in Z, and no descendant of B is in Z.*

If Z blocks every path between two nodes X and Y, then X and Y are d-separated conditional on Z. Therefore they are conditionally independent given Z.

In general, condition on the middle point of a chain or fork will block the path, while conditioning on a collision node will unblock the path. For example, consider the graphical

model in Figure 2.6 panel (a), the only path between X and W is composed of a collider ($X \leftarrow Y \rightarrow Z$) and a fork ($Y \rightarrow Z \leftarrow W$). To block this path, we can either exclude Y (collision node) from the conditioning set or include Z (middle point of the fork) in the set. Thus, X and W are d-separated (and therefore independent) if condition on the empty set, $\{Z\}$, $\{Y, Z\}$, $\{T, Z\}$, and $\{T, Y, Z\}$. In panel (b), an edge between T and W is added, and now there are two paths from X to W. Both paths need to be blocked if X and W are d-separated. Hence, X and W are independent conditional on $\{T\}$, $\{T, Z\}$, $\{T, Y, Z\}$.

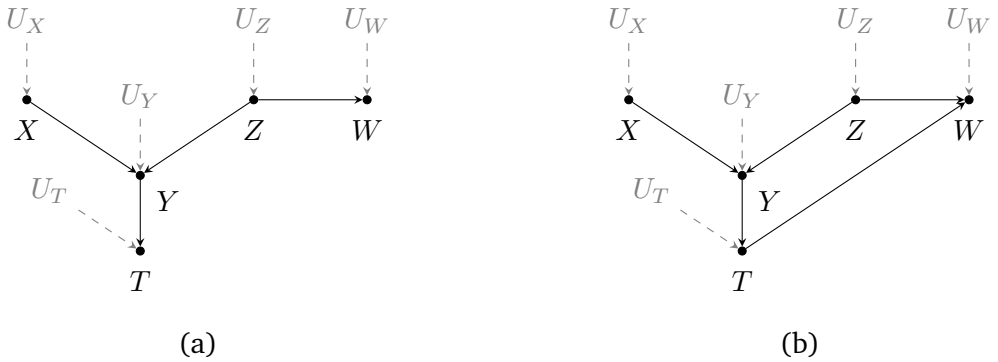


Figure 2.6 Panel (a) and (b) are examples of more complicated causal graphs. Both graphs are combinations of the three fundamental structures.

Definition 4. ***(The Backdoor Criterion)*** *(Pearl et al. 2016) Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.*

From the d-separation criterion, we identified a set of conditioning sets that separate two variables X and Y, i.e. make X and Y conditionally independent. Conditioning sets identified from d-separation need to meet further criteria to satisfy the backdoor criterion. In general, we would like to condition on a set of nodes Z such that (Pearl et al. 2016)

1. We block all spurious paths between X and Y.
2. We leave all directed paths from X to Y unperturbed.
3. We create no new spurious paths.

In practice, we apply the backdoor criterion to obtain the minimal adjustment set needed to estimate the unbiased causal effect from X to Y. We can estimate this unbiased causal effect using regression or probabilistic methods (Cenci & Kealhofer 2022).

2.3 Biases in Regression Models

As we mentioned in the previous section, an SCM is, in theory, falsifiable (Pearl 1995, Cenci & Kealhofer 2022). By testing the conditional independencies entailed in the graphical representation of the SCM, we can validate our causal model on ex-ante data validate our causal model on data. In contrast, regression models are not falsifiable and can only be interpreted and evaluated ex-post.

In structural causal modeling, we explicitly present our hypothesis on the causal mechanisms generating the data, while regression equations merely describe statistical correlations without

making any assumptions about the causal story behind the data. The decisions for control variables in regressions are typically made without clear justification, which leads to various "kitchen sink" regressions[2]. The hope is that regression itself will assign zero to the coefficients of the insignificant variables and attribute the correct sign to all the others. This approach can lead to various biases in estimating the statistical relationships or causal relationships between variables. We introduce two of the bias examples: collider bias and mediator bias.

2.3.1 Collider Bias

Consider the following SCM:

$$X = U_X; \quad Y = 1.5X + U_Y; \quad Z = -Y - 0.75H + U_Z; \quad H = U_H \tag{2.2}$$

where the independent exogenous variables U_X, U_Y, U_Z, U_H are all assumed to be random Gaussian noise (follows $\mathcal{N}(0, 1)$). We generate 800 data samples for X, Y, Z, and H following the structural equations defined in Eq. (2.2). Our goal is to estimate the effect of X on H.

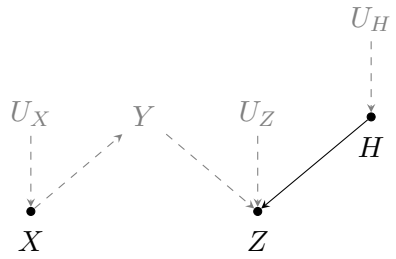


Figure 2.7 Collider bias example. Graphical representation for the SCM stated in Eq. (2.2).

Given the simulated data, we can regress H on X ($H = \alpha X + \epsilon$), and evaluate the coefficient α and its statistical significance. Then we add Z to the regression as a control variable. The results for the two regressions are presented in Table 2.1. We can observe that the coefficient for X is close to 0 and is statistically insignificant (p-value = 0.809 > 0.1) in the first regression model. This follows from the SCM defined in Eq. (2.2) as both X and H are generated from an independent standard normal distribution. In comparison, the second regression model gives us a statistically significant (very small p-value) negative α. The standard errors for the two models do not differ too much, but R^2 is much larger for the second model. This is known as the *collider bias* (Greenland et al. 1999). When we include the middle node of a collider in the causal graph, we will introduce bias in our estimates. By conditioning on Z, X and H become dependent, which would otherwise be independent. This will create a new spurious path from X to H (make them d-connected), which violates the third condition in the backdoor criterion.

Covariates	Coefficient α	Standard error	t-value	p-value	R^2
X	-0.003	0.034	-0.083	0.934	0.000
X, Z	-0.408	0.042	-9.786	0.000	0.203

Table 2.1 The table shows the regression results from data simulated according to Eq. (2.2). The first row shows the result for $H = \alpha X + \epsilon$. The second row shows the result for $H = \alpha X + \gamma Z + \epsilon$.

[2]A regression model that includes a large number of explanatory variables, often without a clear rationale or theoretical justification for their inclusion.

The implication of this section is that we need to select the control set used in regression models carefully. Even in a simple example illustrated above, controlling for the wrong or unnecessary variable can lead to biased results. There are no purely statistical tools that can distinguish between the two models presented in Table 2.1. However, if we use SCM to approach the association estimation, we can construct our causal model and validate it ex-ante through conditional independence tests. We can therefore identify the collision points in the graph through the backdoor criterion and exclude those points from the control set.

2.3.2 Mediator Bias

In SCM, a variable can cause another variable directly or indirectly, through a set of mediating variables. The total causal effect is the sum of the direct and indirect effects. Similarly, we generate 800 data samples from the below SCM with its associated causal graph illustrated in Figure 2.8.

$$X = U_X; \quad Y = 0.5X + U_Y; \quad Z = Y + U_Z; \quad H = Y + 0.75Z + U_H \tag{2.3}$$

where $U_X, U_Y, U_Z, U_H \sim^{i.i.d.} \mathcal{N}(0,1)$. Our goal is still to estimate the effect of X on H. From the causal graph in Figure 2.8, we find that Y is a mediator along the causal path between X and H which blocks all causal paths between X and H. H is a descendent of the mediator Y, for which including Z in the conditioning set will block the path $X \to Y \to Z \to H$ but leave the path $X \to Y \to H$ open.

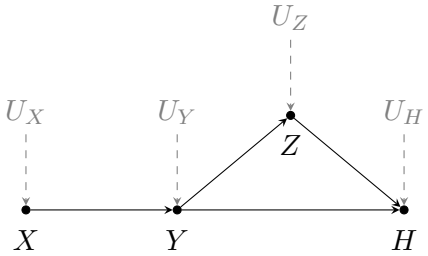


Figure 2.8 Mediator bias example. Graphical representation for the SCM stated in Eq. (2.3).

Covariates	Coefficient α	Standard error	t-value	p-value	R^2
X	0.435	0.052	8.422	0.000	0.082
X, Y	0.067	0.050	1.344	0.179	0.318
X, Z	0.053	0.038	1.374	0.170	0.556

Table 2.2 The table shows the regression results from data simulated according to Eq. (2.3). The first row shows the result for $H = \alpha X + \epsilon$. Coefficient α in this regression is an unbiased estimate of the total effect of X on H. The second row shows the result for $H = \alpha X + \beta Y + \epsilon$, which yields an unbiased estimate of the direct effect of X on H. The third row shows the result for $H = \alpha X + \gamma Z + \epsilon$. α in this regression is a biased estimate of the total effect.

The first row in Table 2.2 shows the estimate for the total effect of X on Z, combining the effect conducted through $X \to Y \to H$ and $X \to Y \to Z \to H$. When controlling for Y in the regression, we measure the direct effect of X on Z as conditioning on Y blocks all the causal paths between X and H. In the third row, we estimate the effect by controlling Z (descendent of the mediator Y). This approach will block the causal effect flows through the path $X \to Y \to Z \to H$, but leave the path $X \to Y \to H$ open. Thus, we will introduce a

bias in the estimate as we remove part of the total effect of X on Z. This is what we call the *mediator bias* and it is a bias of the total effect. Including either Y or Z in the conditioning set will violate the second requirement of the backdoor criterion. Again, the statistical performance (R^2) of the regression model is improved when we add additional control variables.

In conclusion, we apply the backdoor criterion in an SCM to identify the confounders we need to control for to estimate the total causal effect of one variable on another. This approach will avoid biases in the estimate of effect such as collider bias and mediator bias. Through SCM, we also establish a clear hypothesis on the causal story between variables, allowing us to better understand and explain our model. Furthermore, we can validate our model on data by conditional independence testing, which is the topic in the next section.

2.4 Measuring Conditional Independencies

As previously discussed, measuring conditional independence conditions is crucial for structural causal modeling. Through conditional independence testing, we can validate our hypothesis (a graphical causal model) on the causal structure between variables and the underlying data-generating mechanism. In section 2.2, we mentioned two primary limitations of the conditional independence analysis: the inability to test conditional independencies in some causal graphs and the issue of Markov equivalence. Figure 2.5 illustrated three examples in which we have difficulties measuring conditional independencies from the graph.

The case illustrated in Figure 2.5c (identifiability problems due to graph structure alone) can be resolved by counterfactual reasoning (Cenci & Kealhofer 2022). For instance, to prove the causal link between X and Z in Figure 2.5c does exist, we can build a counterfactual hypothesis by removing the edge (X, Z). If our counterfactual hypothesis is true, then the condition $X \perp\!\!\!\perp Z \mid Y$ is true. If the conditional independence test shows X and Z are not independent conditional on Y, we can reject our counterfactual hypothesis and conclude that the causal link between X and Z indeed exists.

The second identifiable issue is the direction of causation, or equivalently, the issue of Markov equivalence. Graphical models (which are directed acyclic graphs or DAGs) in the same Markov equivalence class impose the same set of conditional independencies and therefore are indistinguishable from conditional independencies alone. Markov equivalent DAGs share the same skeleton (same edges if one ignores direction) and the same v-structures (the collider structures), but different directions of edges (Flesch & Lucas 2007). The decision for the direction of causation is often leveraged by the prior knowledge in the domain. Compared to the black-box nature of many statistical approaches, this ambiguity should not be seen as a limitation due to its transparency (Cenci & Kealhofer 2022), allowing researchers to better understand and interpret models rather than just rely on results. It also allows comparison between multiple possible models. By comparing different DAGs, researchers can evaluate the impact of different assumptions on the results and better understand the complexity of causal relationships (Heckman 2008).

So how do we measure the conditional independencies from data? In statistics, X and Y are conditionally independent given Z if and only if $p_{X|Y,Z} = p_{X|Z}$ (or equivalently, $p_{Y|X,Z} = p_{Y|Z}$, or $p_{XY|Z} = p_{X|Z}p_{Y|Z}$). Thus, the most straightforward method to test $X \perp\!\!\!\perp Y \mid Z$ is to estimate the relevant conditional densities and evaluate if the above equations hold. However, in practice, conditional density estimation in high dimensions is challenging, due to the curse of dimensionality (exponential increase in data volume associated with the number of data dimensions). Fortunately, there have been analytical solutions for conditional independence tests in linear systems, yet SCM does not assume linear relationships between variables.

Several regression-based or model-free methods have been developed to estimate conditional independence in the presence of nonlinearity. We propose to use the model-free kernel conditional independence test (KCIT) developed in Zhang, Peters, Janzing & Schölkopf (2012) and Strobl, Zhang & Visweswaran (2019).

An alternative characterization of conditional independence is given in terms of the cross-covariance operator on Reproducing Kernel Hilbert Space (RKHS) (Fukumizu et al. 2004). For the random vector (X, Y) on domain $\mathcal{X} \times \mathcal{Y}$, the cross-covariance operator Σ_{XY} from $\mathcal{H}_\mathcal{X}$ to $\mathcal{H}_\mathcal{Y}$[3] is defined as follow:

$$\langle f, \Sigma_{XY} g \rangle = \mathbb{E}_{XY}[f(X)g(Y)] - \mathbb{E}_X[f(X)]\mathbb{E}_Y[g(Y)], \quad \forall f \in \mathcal{H}_\mathcal{X}, g \in \mathcal{H}_\mathcal{Y} \tag{2.4}$$

Consequently, we can define the partial cross-covariance operator of (X, Y) given Z as:

$$\Sigma_{XY \cdot Z} = \Sigma_{XY} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY} \tag{2.5}$$

which can be intuitively interpreted as the partial covariance between $f(X), \forall f \in \mathcal{H}_\mathcal{X}$ and $g(Y), \forall g \in \mathcal{H}_\mathcal{Y}$ given $h(Z), \forall h \in \mathcal{H}_\mathcal{Z}$ (Zhang et al. 2012). Fukumizu, Gretton, Sun & Schölkopf (2007) proposed a proposition that, under loose conditions[4],

$$X \perp\!\!\!\perp Y \mid Z \iff \Sigma_{\ddot{X}Y \cdot Z} = 0, \quad \ddot{X} = (X, Z) \tag{2.6}$$

Now we consider the following hypothesis for conditional independence testing:

$$\begin{aligned} H_0 &: X \perp\!\!\!\perp Y \mid Z \\ H_1 &: X \not\perp\!\!\!\perp Y \mid Z \end{aligned} \tag{2.7}$$

Since $\Sigma_{\ddot{X}Y \cdot Z} = 0 \iff \|\Sigma_{\ddot{X}Y \cdot Z}\|_{HS}^2 = 0$, using the characterization of conditional independence in equation (2.4), the above hypothesis is equivalent to

$$\begin{aligned} H_0 &: \|\Sigma_{\ddot{X}Y \cdot Z}\|_{HS}^2 = 0 \\ H_1 &: \|\Sigma_{\ddot{X}Y \cdot Z}\|_{HS}^2 > 0 \end{aligned} \tag{2.8}$$

where $\|\cdot\|_{HS}^2$ is the squared Hilbert-Schmidt norm in Euclidean space. The kernel conditional independence test uses an empirical estimate $\mathcal{S} = n\|\Sigma_{\ddot{X}Y \cdot Z}\|_{\widehat{HS}}^2$ for the squared Hilbert-Schmidt norm of the partial cross-covariance operator, which can be computed using centralized kernel matrices (see Theorem 4 and Proposition 5 of Zhang et al. (2012) for more details). $\mathcal{S}$ is a test statistic to determine whether to reject H_1 in hypothesis 2.8.

An implementation issue for KCIT is the test scales cubically with sample size. To deal with conditional independence tests for large datasets, Strobl, Zhang & Visweswaran (2019) derived two relaxations: the randomized conditional independence test (RCIT) and the randomized conditional correlation test (RCoT) which both approximate KCIT by utilizing random Fourier features. The work proves that the squared Hilbert-Schmidt norm of the empirical partial cross-covariance matrix can be replaced by a squared Frobenius norm $\mathcal{S} = n\|\Sigma_{\ddot{X}Y \cdot Z}\|_{\widehat{F}}^2$, and such approximation leads to efficient implementation. In fact, both RCIT and RCoT scale linearly with sample size in practice. We propose to use RCIT in our research[5].

[3] $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$ are the corresponding RKHS for $\mathcal{X}$ and $\mathcal{Y}$.

[4] The details of RKHS and related theorems are beyond the scope of this report, one can refer to Fukumizu et al. (2004) and Fukumizu et al. (2007) if interested in.

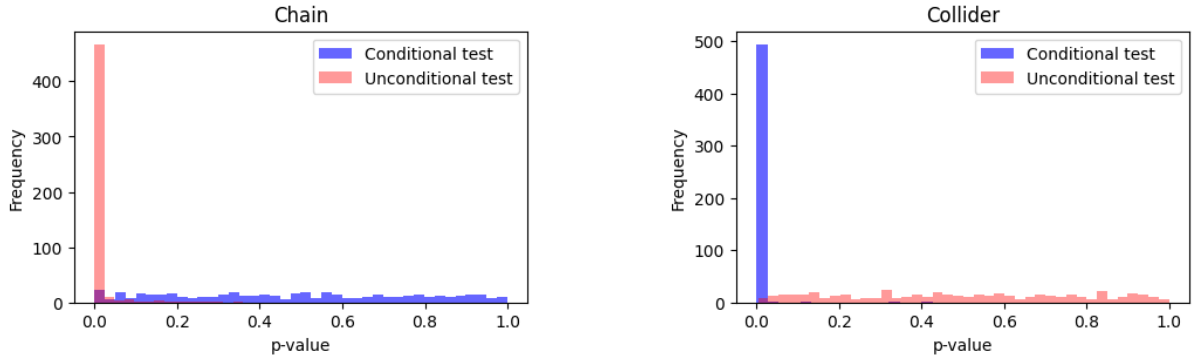
[5] An implementation of the RCIT algorithm is available at `https://github.com/ericstrobl/RCIT`

For an illustrative example of how we test conditional independencies in a causal graph, consider a chain (in Figure 2.2) and a collider (in Figure 2.4). We assume simple structural equations for the chain and collider as follows:

$$Chain: X = U_X; \quad Y = 2X + U_Y; \quad Z = \cos(Y) + U_Z \tag{2.9a}$$

$$Collider: X' = U'_X; \quad Y' = U'_Y; \quad Z' = \sin(X') + 0.5Y' + U'_Z \tag{2.9b}$$

The independent exogenous variables $U_X, U_Y, U_Z, U'_X, U'_Y, U'_Z$ are all assumed to be random Gaussian noise. We generate 1000 random i.i.d. samples for variables X, Y, Z, X', Y', and Z' from their corresponding SCM.



(a) Chain **(b)** Collider

Figure 2.9 Histograms of the p values of conditional and unconditional independence tests for a chain and a collider. We simulated data for each causal graph using the SCM defined in equations (2.9a) and (2.9b). Panel (a) shows that $X \perp\!\!\!\perp Z \mid Y$ and X dependent to Z for the chain structure. Panel (b) shows $X \perp\!\!\!\perp Z$ and X dependent to Z conditional on Y for the collider structure.

For a chain structure in Figure 2.2, we know that X is conditionally independent of Z given Y. However, testing such a condition alone is not sufficient is not enough to confirm that the graph is a chain, as the same conditional independence could also be met by a graph where the variables have no causal relationships (Cenci & Kealhofer 2022). To validate our simulated data following structural equations in (2.9a) support the chain structure, we need to test: (1) $X \perp\!\!\!\perp Z \mid Y$ is true, and (2) $X \perp\!\!\!\perp Z$ is not true. In other words, we need to show that X and Z are dependent but become conditionally independent once we condition on Y, i.e. the dependence between X and Z is caused by the confounding variable Y.

Verifying these two conditions also eliminates the possibility of other structures[6]. For instance, if the true causal graph is a collider in Figure 2.4, then X' and Z' are unconditionally independent but are dependent if we condition on the collision point Y'. In this case, the conditional independence test results should be (1) $X' \perp\!\!\!\perp Z' \mid Y'$ not true, and (2) $X' \perp\!\!\!\perp Z'$ true. In Figure 2.9, we simulate a chain and a collider and use the RCIT algorithm to test for conditional and unconditional independencies. This process is repeated for 500 times. Figure 2.9 shows that for a chain structure, the p values for conditional independence tests are uniformly distributed between 0 and 1, while the unconditional independence test results concentrate around 0. The test results are the opposite for a collider structure as illustrated in panel (b). We use $\mathcal{X}^2$-test to assess the uniformity in p values for this research. A p-value smaller than 0.1 from $\mathcal{X}^2$-test denotes the p values distribution of repeated conditional (unconditional) independence tests largely deviates from a uniform distribution. In this case, we should reject the

[6]We cannot rule out the possibility of a fork due to the identifiability issue discussed before. To deal with such ambiguity, we usually choose directions of causation using prior domain knowledge in empirical research.

null hypothesis in the conditional (unconditional) independence test, i.e. the two variables are conditional (unconditional) independent, and revise our causal model.

Finally, we note an important limitation of conditional independence tests when determinism is included in the model. As we will encounter later in the empirical causal analysis in chapter 4, carbon intensity is the ratio of emissions and sales and is totally determined by these two variables. This determinism means that once emissions and sales are known, there is no residual uncertainty about carbon intensity. Thus, statistical testing of these conditional independence conditions is not applicable, as they are inherently satisfied by the deterministic nature of the relationship. In general, determinism in a causal graph violates the assumption of *faithfulness*: the joint distribution of a process X with graph G satisfies the *faithfulness* condition if and only if, for all disjoint subsets of nodes (or single nodes) A, B, $C \in G$, conditional independence implies separation in the graph (Runge 2018, Cenci & Kealhofer 2022). Determinism can be addressed explicitly under certain conditions (Daniusis et al. 2012, Janzing et al. 2012). This topic is beyond the scope of this report.

Chapter 3

Data and Baseline Regression

In this section, we will discuss the data we are using, and perform the baseline regression following a recent study in empirical finance (Bolton & Kacperczyk 2021). The cross-sectional regression addresses the statistical correlations between different CO_2 emission variables and stock returns.

3.1 Data

Our database covers the time period between 2011 and 2020 in the US market. We have sourced emission data from TruCost[1], stock returns from Refinitiv, and corporate fundamentals from COMPUSTAT. The intersection of the three datasets yields 2,293 unique firms; after deleting firms with missing emission variables or cross-sectional return variables, we obtain 1,423 unique firms corresponding to balanced samples of 81,605 firm-month observations[2]. The summary statistics for all variables in our final sample is outlined in Table 3.1. We compute the variables following a similar measure to Bolton & Kacperczyk (2021).

3.1.1 Emission Data

TruCost follows the Greenhouse Gas (GHG) Protocol which sets the standards for corporate emission measurement. The emission data are measured in tons of CO_2 equivalent (tCO_2e) per year, and the GHG protocol distinguishes between three sources of emissions. Scope 1 emissions reflect direct emissions from the establishment owned or controlled by a company. For example, Scope 1 emissions include the emissions produced by a factory's onsite natural gas boiler used for heating. Scope 2 emissions are from purchased heat, steam, electricity, or other sources of energy consumed by the company. Scope 3 emissions are the indirect emissions from operations and products that are not owned or controlled by the firm; these include the emissions from the supply chain of a company and the usage of the company's products by consumers. Scope 1, and 2 emissions are easier to measure and are reported more systematically and estimated more accurately (Bolton & Kacperczyk 2021, Busch et al. 2018). In contrast, scope 3 emissions are

[1] TruCost compiles its data from a variety of publicly available sources, including company financial reports, environmental data sources, and information published on company websites and other public platforms. If a firm does not disclose emissions data voluntarily, its environmental impacts (or emission data) are estimated using an environmentally extended input-output (EEIO) model by TruCost. This model combines industry-specific ecological impact data with quantitative macroeconomic data on the flow of goods and services between different economic sectors. (Aswani et al. 2024).

[2] Our initial database is much smaller, compared to Bolton & Kacperczyk (2021) which containing approximately 189,000 year-month observations. This might influence subsequent results.

difficult to measure and estimate due to their complexity and indirect nature, resulting in low accuracy in scope 3 emission measurements (IBM team 2023). Therefore, emissions in Scope 3 will not be considered in this project, and only Scope 1 and 2 emissions are used. We report the summary statistics of the emission variables in Panel A of Table 3.1.

Table 3.1 Summary statistics for emission variables and cross-sectional return variables

This table reports summary statistics (averages, medians, and standard deviations) for the variables used in our analysis. Panel A reports the emission variables. Panel B reports the cross-sectional return variables. Following Bolton & Kacperczyk (2021), different variables have been winsorized at the specified percentages to mitigate the influence of outliers.

Variable	Winsorization [cutoff (%)]	Mean	Median	Std. Dev.
Panel A: Emission variables				
Log (Carbon Emissions Scope 1)	-	9.88	9.82	2.98
Log (Carbon Emissions Scope 2)	-	10.10	10.30	2.46
Growth Rate in Carbon Emissions Scope 1	2.5	0.06	0.02	0.30
Growth Rate in Carbon Emissions Scope 2	2.5	0.07	0.02	0.29
Carbon Intensity Scope 1/100	2.5	1.16	0.13	3.59
Carbon Intensity Scope 2/100	2.5	0.30	0.17	0.37
Panel B: Cross-sectional return variables				
RET (%)	-	1.01	0.74	12.47
LOGSIZE	-	8.01	8.00	1.81
B/M	2.5	0.59	0.44	0.54
LEVERAGE	2.5	0.28	0.25	0.21
MOM	0.5	0.15	0.08	0.50
INVEST/A	2.5	0.03	0.02	0.04
ROE	2.5	6.71	9.51	31.58
LOGPPE	-	6.75	6.83	2.24
BETA	0.5	1.11	1.05	0.72
VOLAT	0.5	0.11	0.09	0.06
SALESGR	0.5	0.02	0.02	0.31
EPSGR	0.5	0.01	0.00	0.23

We use three emission variables to describe the emission performances of firms, including total emissions, annual emission growth rate, and emissions intensity. Firms' total emissions straightforwardly reflect their emissions levels. In our sample, the average scope 1 emission is 1.36 million tons, which is relatively higher than the average scope 2 emission, at 260,000 tons of CO_2 equivalent. The scope 1, 2 emissions measured in tCO_2e are normalized using the natural log scale to manage the range of emission values and reduce data skewness. We also present the year-to-year growth rate (in percentage) for each emission scope. The Carbon intensity is the ratio of emissions to net sales, measured in tons of CO_2 equivalent per million US dollars of company sales. By scaling with sales, emission intensity avoids the mechanical correlations between emissions and firm size, which could lead to a better understanding of a firm's emission performance (Aswani et al. 2024). Emission growth rates and Carbon intensities are winsorized at 2.5% to neutralize the impact of outliers.

3.1.2 Cross-sectional Return Variables

We define the cross-sectional return variables following Bolton & Kacperczyk (2021). In the baseline regression described in equation 3.1, the dependent variable $RET_{i,t}$ is the monthly stock return for firm i in month t. Returns greater than 100% are removed from the dataset to manage the outliers. For the host of firm-specific variables in the control set[3], $LOGSIZE_{i,t}$ is the natural logarithm of firm i's market capitalization at the end of year t in million dollars; $B/M_{i,t}$ is computed as the book value of equity divided by its market value of equity for firm i; $LEVERAGE_{i,t}$ is the book value of leverage (book value of debt divided by the book value of assets) in firm i; $MOM_{i,t}$ is the average of the most recent 12 months' stock returns on firm i, leading up to and including month $t-1$; $INVEST/A_{i,t}$ indicates firm i's capital expenditure (CAPEX) divided by its book value of assets; $ROE_{i,t}$ is firm i's return on equity, computed as the ratio of net income divided by shareholder's equity; $LOGPPE_{i,t}$ represents the natural logarithm of plant, property, and equipment (PPE) of a firm in million dollars; $BETA_{i,t}$ denotes the market beta calculated over a one-year period using the Capital Asset Pricing Model (CAPM) for firm i in year t; $VOLAT_{i,t}$ is the standard deviation of returns based on stock returns in the past 12 months; $SALEGR_{i,t}$ measures the year-to-year change in revenues, normalized by the firm's market capitalization; $EPSGR_{i,t}$ stands for the annual change in earnings per share (EPS), normalized by the firm's equity price. The summary statistics is presented in Table 3.1 Panel B.

3.2 Baseline Regression

From the data preparation section, we obtain three categories of emissions: total emissions, year-to-year growth rate in emissions, and emission intensities. We now examine the relationship between firms' Carbon emissions and their corresponding stock returns in the cross-section. Following the regression in Bolton & Kacperczyk (2021, p. 530), the baseline regression model is as follows:

$$RET_{i,t} = a_0 + a_1 Emissions_{i,t} + a_2 Controls_{i,t-1} + \mu_t + \sigma_{industry} + \epsilon_{i,t} \tag{3.1a}$$

$$\begin{aligned} Controls_{i,t-1} = (&LOGSIZE_{i,t-1},\ B/M_{i,t-1},\ LEVERAGE_{i,t-1},\ MOM_{i,t-1},\\ &INVEST/A_{i,t-1}, ROE_{i,t-1},\ LOGPPE_{i,t-1},\ BETA_{i,t-1},\\ &VOLAT_{i,t-1},\ SALEGR_{i,t-1},\ EPSGR_{i,t-1}) \end{aligned} \tag{3.1b}$$

where the dependent variable $RET_{i,t}$ represents the monthly stock return for firm i in month t, and $Emissions_{i,t}$ refers to one of the emission variables: natural logarithm of total firm-level emissions, the year-to-year emissions growth rate, and the Carbon intensities in Scope 1 and Scope 2 (as shown in Table 3.1 Panel A)[4]. The control variables are illustrated in equation (3.1b), including 11 firm-specific variables of firm i in month $t-1$ potentially influencing stock returns in month t. Full definitions for the control variables are given in section 3.1.2. The coefficients μ_t and $\sigma_{industry}$ represent the fixed effects for year-month and industry, respectively. We estimate the regression model using pooled OLS and the results for regression are reported in Table 3.2. We are mainly interested in the coefficient a_1 which reflects the estimate for associations between firms' emissions and their stock returns. Note that we do not directly compare our results to the results in Bolton & Kacperczyk (2021) due to dataset limitations.

[3] Compared to Bolton & Kacperczyk (2021), our analysis lacks the variable *HHI* in the control set due to insufficient data in our original dataset. This omission may contribute to differences in subsequent results.

[4] Emission variables and LOGSIZE, B/M, LEVERAGE, INVEST/A, ROE, SALEGR, EPSGR in cross-sectional return variables are obtained as annual data. We assume that the annual variables remain the same within the same year. Hence, for instance, the variable $Emissions_{i,t}$ describes firm i's emission in month t.

In Table 3.2 Panel A, we report the estimated coefficients for the natural logarithm of total emissions in Scope 1 and 2. Scope 1 total emissions have a variable relationship with stock returns depending on the fixed effects included. The association is positive and significant when only year-month effects are included but becomes insignificant when both year-month and industry effects are considered. Scope 2 total emissions consistently show a highly significant negative relationship with stock returns across all specifications.

For emission growth rates (Table 3.2 Panel B), both Scope 1 and Scope 2 emissions do not show any significant relationship with stock returns across all specifications.

For carbon intensities (Table 3.2, Panel C), Scope 1 carbon intensity has a significant positive relationship with stock returns when no fixed effects or only year-month fixed effects are included. However, this relationship becomes insignificant when both year-month and industry-fixed effects are considered. Scope 2 carbon intensity does not show a significant association with stock returns in any specification.

In conclusion, Scope 1 total emissions and Scope 1 emission intensity show a positive association with stock returns, but the statistical significance of the associations depends on the inclusion of fixed effects. Scope 2 total emissions tend to have a more consistently significant negative relationship with stock returns compared to Scope 1 emissions and intensity measures. The lack of significance in emission growth rates may suggest that the level of emissions might be more relevant for stock returns than their growth rates.

Table 3.2 Baseline regression results

The table shows the OLS coefficients for emission variables in the baseline regression defined in Eq. 3.1. We report the statistical significance p values in the parentheses below the coefficients. The dependent variable is RET. The main independent variable CO_2 emissions takes one of the following forms: log total emission, emission growth rate, and emission intensity for both Scope 1 and Scope 2. All variables are defined in Table 3.1. Columns (1) and (2) show the regression coefficients without considering year-month or industry fixed effects, columns (3) and (4) additionally include the year-month effects, and columns (5) and (6) take all fixed effects into account. Panel A reports the results for the natural logarithm of total emissions; Panel B reports the results for the growth rate in emissions; Panel C reports the results for carbon emission intensity. ***, **, * indicates statistical significance at 1%, 5%, and 10%, respectively.

Panel A: Total Emissions						
Variables	(1) RET	(2) RET	(3) RET	(4) RET	(5) RET	(6) RET
Log (Scope 1)	0.0512*		0.5601***		0.1401	
	(0.074)		(0.000)		(0.240)	
Log (Scope 2)		-1.6088***		-0.7432***		-0.8454***
		(0.000)		(0.000)		(0.000)
Observations	81605	81605	81605	81605	81605	81605
Year-month F. E.	No	No	Yes	Yes	Yes	Yes
Industry F. E.	No	No	No	No	Yes	Yes
Panel B: Emission Growth Rates						
Growth Scope1	0.0802		0.0507		0.0738	
	(0.582)		(0.722)		(0.604)	
Growth Scope2		-0.0114		-0.0152		-0.0105

(continued on next page)

(continued)

Variables	(1) RET	(2) RET	(3) RET	(4) RET	(5) RET	(6) RET
		(0.939)		(0.917)		(0.942)
Observations	81605	81605	81605	81605	81605	81605
Year-month F. E.	No	No	Yes	Yes	Yes	Yes
Industry F. E.	No	No	No	No	Yes	Yes
			Panel B: Carbon Intensity			
Intensity Scope1	0.0349***		0.0303**		0.0017	
	(0.006)		(0.015)		(0.914)	
Intensity Scope2		-0.1313		-0.1223		-0.1309
		(0.280)		(0.304)		(0.359)
Observations	81605	81605	81605	81605	81605	81605
Year-month F. E.	No	No	Yes	Yes	Yes	Yes
Industry F. E.	No	No	No	No	Yes	Yes

Chapter 4

Methods

As introduced in section 2.3, selecting the right control variables is crucial for the success of a regression model. Adopting control variables that are spuriously correlated can increase model performance, but can also introduce biases in the estimates of effects, leading to misleading conclusions that appear to be supported by observational data but are not reflective of the true data-generating mechanism. We aim to improve the cross-sectional regression model described in Eq. (3.1) by systematically identifying the necessary control set through the construction of causal structures using structural causal modeling (SCM).

In this chapter, we address the issues of identifying causal structures among the returns, emission variables, and other firm-level financial variables. This involves establishing assumptions on the causal relationships between the variables of interest, for instance, how a specific firm characteristic, such as firm size or leverage, influences its stock returns. A popular framework addressing the identification issues is structural causal modeling, which we will adopt in the following research.

By establishing our hypotheses on variables' causal structures using SCM, we present assumptions on the data-generating mechanism driven by the causal story behind the variables. Each SCM can be uniquely represented by an associated causal graph (Pearl et al. 2016) with well-defined directions of causation. For instance, $X \to Y$ in a graph indicating X is a direct cause of Y. According to the theory of Bayesian networks, causal graphs, which are directed acyclic graphs (DAGs), imply a set of conditional independencies that can be tested on empirical data (Pearl 1995). Hence SCM is ex-ante falsifiable, in contrast to the standard regression model, which can only be interpreted and evaluated ex-post (Cenci & Kealhofer 2022). By examining these causal links through conditional independence tests, we correct and validate our hypotheses on causal structures.

In summary, the SCM approach includes 4 steps: (1) formulate assumptions on causal relationships between variables and present the hypotheses by a causal graph; (2) estimate the conditional independencies implied by the graph to validate our hypotheses; (3) use the back-door criterion to select the necessary control variables; (4) re-estimate the associations between emissions and stock returns. A more detailed introduction to SCM and conditional independence tests can be found in chapter 2.

4.1 Structual Causal Model Development

In this section, we formulate assumptions on the causal structure for stock returns, emissions, and firm characteristics using the structural causal model. As introduced in chapter 2, an SCM is a set of structural equations with well-defined directions indicating the causal relationships

between variables(Pearl 2009). In our approach, we make minimal assumptions about the exact analytical form of the equations and focus on the qualitative causal relations entailed in the graphical representation of SCM, the causal graph.

A causal graph is a directed acyclic graph (DAG) where nodes represent variables, and directed edges represent causal influences between these variables. If X is a parent or ancestor of Y, then X is a direct cause or cause of Y. Each causal graph entails a set of conditional independencies that can be tested ex-ante on the data. Thus, a SCM can be, in principle, falsified (Cenci & Kealhofer 2022). We will validate our model developed in this section using conditional independence tests in section 4.2. As introduced in chapter 2, every endogenous variable has at least one unobserved exogenous variable influencing it. In the empirical analysis, we only present unobserved exogenous variables linked to multiple endogenous variables for simplicity. The neglected exogenous variables are perceived as idiosyncratic shocks primarily affecting only one endogenous variable and are usually treated as random noise. We include the fixed-time effect in conditional independence tests (section 4.2) to control for systemic temporal variations, such as changes in macroeconomic conditions, policy shifts, or global events.

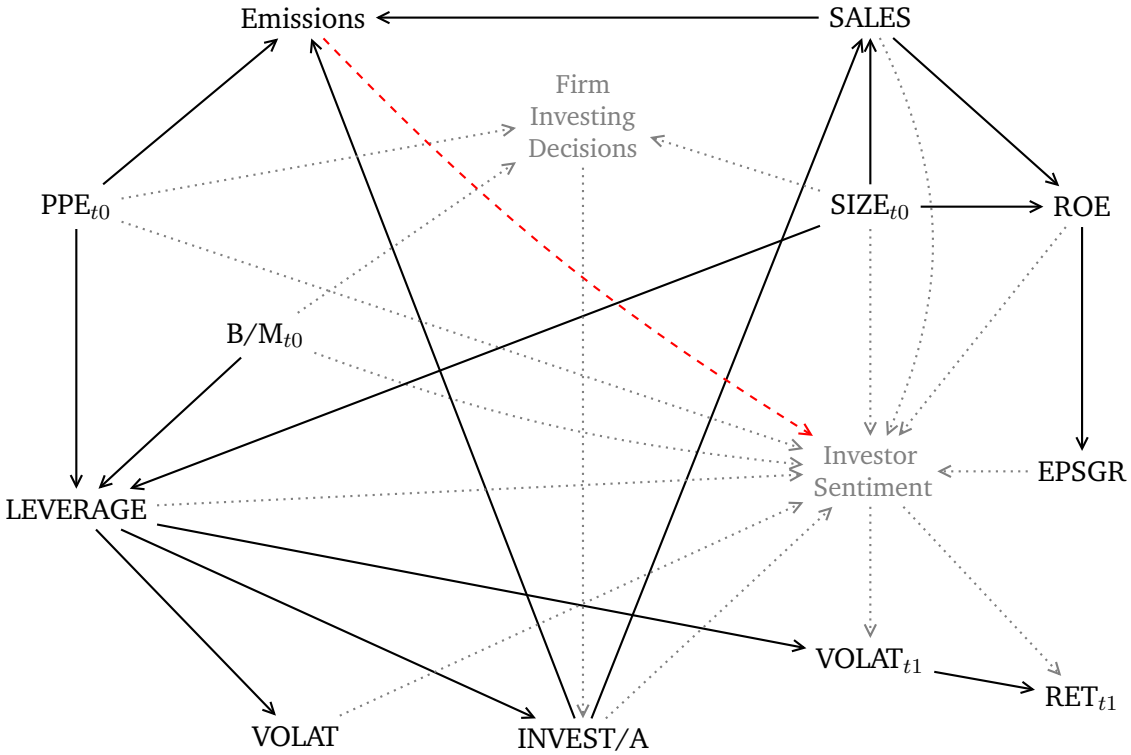


Figure 4.1 SCM for carbon emissions and stock returns. Firm investing decisions and investor sentiment are two unobserved variables and are painted in gray. The directed edges connected to those two unobserved mediators are drawn as dotted gray lines. Time t_0 is the initial time, t_1 is the final time, and variables without a time mark are considered to fall within the interval between t_0 and t_1. The unobserved edge from *Emissions* to *Investor Sentiment* marked in red dashed line is the essential unobserved link we are interested in.

Note that, for illustrative purposes, we directly use variable names as nodes in our causal diagrams. Whereas in chapter 2, we represented nodes visually and labeled them with variable names beside the nodes.

Our hypothesis on causal structure for stock returns and emissions is presented in Figure

4.1. We derive our model by looking at the two main factors driving emissions and stock returns: firm investing decisions and investor sentiment. Firm investing decisions (eg. investing in green technology or increasing the production scale) are made internally based on various firm-specific variables. Investor sentiment in the market reflects how investors perceive firms' characteristics, which in turn influences their stock trading decisions and ultimately determines stock returns. These two factors are presented as unobserved variables in gray in Figure 4.1. We are essentially interested in how the investors perceive firm emissions, i.e. do investors care about carbon emissions? In our structural causal model, this question is translated to the existence of the causal link from emissions to market (the red dotted link).

As shown in Figure 4.1, PPE_{t_0}, $SIZE_{t_0}$, and B/M_{t_0} are initial tangible assets, firm size (computed as the market capitalization), and book-to-market ratio for a firm. They are assumed to be fundamental characteristics of the firm at the initial time t_0 and are presented as root variables in the graph[1]. $VOLAT_{t_1}$ and RET_{t_1} are volatility reflecting the fluctuation in stork returns in percentage and stock returns at time t_1. The temporal structure of other firm variables including *SALES*, *INVEST/A*, *LEVERAGE*, *VOLAT*, *ROE*, and *EPSGR* is under-specified on the graph and is assumed to operate within the interval between t_0 and t_1. *SALES* represents the firm's sales revenue, *INVEST/A* is the investment-to-asset ratio, *LEVERAGE* is the debt-to-equity ratio, *VOLAT* denotes the stock prices volatility, *ROE* stands for return on equity (computed as the ratio between net income and shareholders' equity), and *EPSGR* indicates earnings per share growth rate.

A firm is assumed to make its investing decisions based on its initial size ($SIZE_{t_0}$), amount of tangible assets (PPE_{t_0}), and book-to-market ratio (B/M_{t_0}). Size and tangible assets indicate the investment capacity of a firm. A higher B/M ratio suggests that the firm is undervalued by the market, potentially indicating strong growth opportunities. Firms with a high B/M ratio might be motivated to invest aggressively to capitalize on these opportunities and improve their market valuation (Zhang 2005). Other factors potentially influencing firm investing decisions, like sales at t_0, are assumed to be reflected in the initial size, tangible assets, or B/M ratio (through the impact on firm valuation in the market), and thus are not explicitly included in the causal graph for simplicity.

Firm investing decisions drive changes in investment, which can be reflected in the investment-to-asset ratio. Leverage also drives *INVEST/A* by providing firms with additional capital to invest in growth and expansion projects. This increased access to funds enables leveraged firms to make substantial investments relative to their existing asset base. A firm may decide to use its investment in green technologies and efficiency improvements, which explains the causal link from *INVEST/A* to *Emissions*. Alternatively, the investing decisions can drive sales growth by expanding production capacity (*INVEST/A* $\rightarrow$ *SALES*). On the one hand, sales are closely related to carbon emissions. Higher sales typically result in higher production levels and increased operational activities, which leads to higher emission levels. On the other hand, rising sales directly boost net income, thereby driving *ROE* (net income divided by equity). A strong return on equity indicates that the company is effectively using its equity base to generate income. This efficiency often results in substantial profits, which can be distributed to shareholders (Penman 2013).

As discussed in Titman & Wessels (1988) and Rajan & Zingales (1996), firms with larger sizes and more tangible assets have easier access to debt financing and thus tend to have higher leverage. Additionally, companies with a high B/M ratio (i.e., relatively low market valuations) prefer to use debt financing rather than issuing equity to avoid diluting shareholders at low valuations (Fama & French 1995). Thus, we also assume $SIZE_{t_0}$, PPE_{t_0}, and B/M_{t_0} to be the direct cause of leverage. High leverage means a high debt burden which can lead to increasing

[1] Exogenous variables that influence PPEt_0, SIZEt_0, and B/Mt_0 are neglected from the graph for simplicity.

volatility in the firm's financial performance.

Finally, we look at causal links related to the two core variables of interest: emissions and stock returns. In our hypothesis, a firm's emission level is directly driven by sales, amount of tangible assets, and investment. Firm characteristics including size and book-to-market ratio also indirectly influence emissions through firm investing decisions or sales. We assume investor sentiment to be driven by all firm characteristics, including emissions and firm investing decisions. Investor perception affects Investor behavior and market demand for stocks, which in turn determines stock returns. If investors do care about carbon emissions, this concern can lead to greater uncertainty about the company's future financial performance. Thus, we also include $VOLAT_{t_1}$ in the model.

We use three measures to evaluate the pollution level of a firm: total emissions, emission growth rate, and carbon intensity. The first two variables are determined by the emission level of a firm and hence can be substituted into the *Emission* node in Figure 4.1. In contrast, carbon intensity, computed as the ratio of emissions to sales, is determined by both the emission level and sales. Therefore, we add links *Emissions* $\rightarrow$ *Intensity* and *SALES* $\rightarrow$ *Intensity* to the original causal graph. Note that emission intensity is entirely determined by emissions and sales, and there is no (external) exogenous variable that causes it. According to the definition[2] of SCM in chapter 2, emission intensity is not a rigorous endogenous variable. Therefore, we include intensity in Figure 4.2 as a derived quantity completely determined by other variables, and denote the edge connected to it differently in gray.

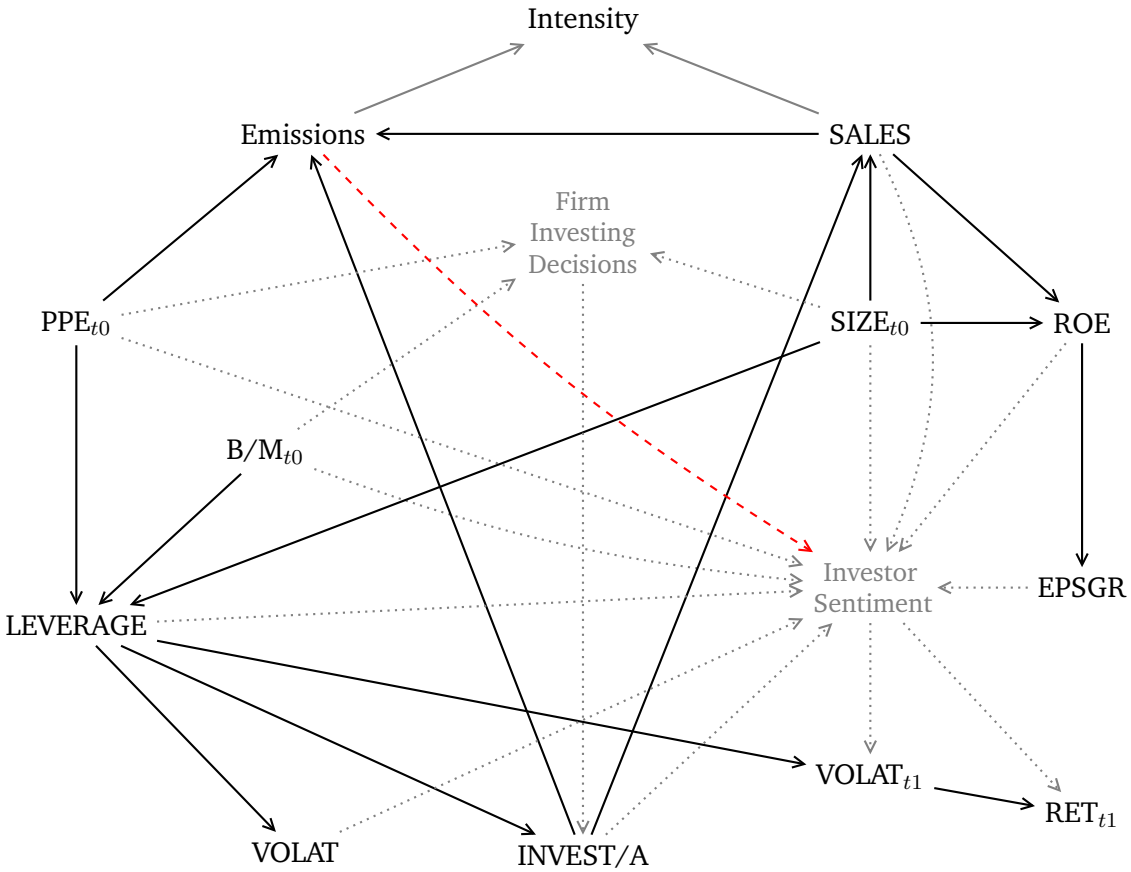


Figure 4.2 SCM for carbon emissions and stock returns. We add carbon intensity to this graph. Note that emission intensity is not a rigorous endogenous variable, but a derived quantity totally determined by sales and emissions. The edge linked to Intensity is marked differently in gray.

[2]Every endogenous variable in a model is a descendent of at least one exogenous variable.

Aligning with the baseline regression illustrated in section 3.2, we wonder about the relationship between carbon emissions and stock returns. Observing from our causal hypothesis illustrated in Figure 4.1, there are two causal paths from emissions to stock return: *Emissions* $\rightarrow$ *Market* $\rightarrow$ *RET*$_{t_1}$ or *Emissions* $\rightarrow$ *Market* $\rightarrow$ *VOLAT*$_{t_1}$ $\rightarrow$ *RET*$_{t_1}$. In other words, if the unobserved red link in the proposed causal graphs (Figure 4.1) indeed exists, then investors do care about firm emissions, and this investor perception influences stock returns directly or through volatility. As we will show in the next section 4.2.3, this link will be tested on empirical data in a counterfactual approach.

Note that in the baseline regression following the study by Bolton & Kacperczyk (2021), emissions and stock returns at the same time t are used. However, in our structural causal model, we found that to investigate the question, *Do investors care about carbon emissions?*, it is more causally reasonable to start with emissions from the previous time period to stock returns. This allows time for investors to receive information and react to a firm's emissions, which in turn affects stock returns. Therefore, we measure the effect of Emissions on RET$_{t_1}$.

Compared to the baseline regression, we exclude two firm variables in our causal model, Beta and momentum (MOM). They are included in the baseline control set to isolate the effect of emissions on stock returns and are known to predict returns (Fama & MacBeth 1973, Carhart 1997). Beta measures a stock's systemic risk relative to the overall market, reflecting the stock's sensitivity to market fluctuations. In our model, *Investor Sentiment* already captures investor decisions and market sentiment, so the path *VOLAT* $\rightarrow$ *Investor Sentiment* $\rightarrow$ *RET*$_{t_1}$ should encompass the role of Beta. Thus, including Beta separately in the SCM would be redundant. MOM represents the cumulative stock return in the past. Our SCM determines return at time t_1 by integrating emissions, firm fundamentals, and market factors. Similarly, returns in the past are inherently reflected through these integrated factors. Therefore, we exclude MOM from the model to avoid redundancy and maintain simplicity.

Overall, we believe that the framework presented in Figure 4.2 provides a reasonable and well-explained causal story behind emissions and stock returns. Our causal graph is a valid DAG and contains no cycle. Note that our model assumes that all changes occur within the time interval between t_0 and t_1, i.e. firms and market make investing decisions based on firm data in this period[3]. The time interval can be adjusted according to the available time period in the dataset, eg. for a monthly dataset, this interval is likely to be two months; for a yearly dataset, the interval can be two years. However, the true temporal dynamics of how firm decisions and market reactions unfold might be more complex and extend beyond this interval. We will further discuss the limitations of our SCM in chapter 6.

4.2 Conditional Independence Tests

Now we have proposed a hypothesis on the causal relationships between emissions, stock returns, and other firm characteristics through a structural causal model in 4.2. In this section, we perform conditional independence tests to validate our hypothesis.

To test the existence of a causal path between two variables *X* and *Y* on the causal graph, we need to test the conditional independencies between them. If the graph does not imply any conditional independence between *X* and *Y* due to the graph structure, we test the counterfactual hypothesis: the causal graph without the link we assumed exists. If the conditional independence entailed in the counterfactual hypothesis is supported by data, then the link between X

[3]It is also assumed that investors in the market have access to relevant firm variables in time, allowing them to make informed decisions based on up-to-date information. This may not always be the case in reality due to information asymmetry, delays in data reporting, and the varying speed at which information is disseminated and processed by the market. However, this is beyond the scope of this report.

and Y should be removed; if not supported by data, then the graph without the assumed link is false and the link exists.

As introduced in section 2.4, the conditional independencies entailed in our graphical causal model (Figure 4.1) can be identified by applying *d-separation*. If *X* and *Y* are unconditionally dependent, there could be a causal path between them. However, if *X* and *Y* become independent conditional on *Z*, and Z is not on the causal path between *X* and *Y*, then the dependence between X and Y is likely due to the confounding variables in Z. Their association becomes spurious and we reject the existence of a causal path between *X* and *Y*.

Knowing what conditional independence condition to test on, the final question is how do we assess if $X \perp\!\!\!\perp Y \mid Z$ from data. The general procedure is introduced in section 2.4. We use KCIT as it makes minimal assumptions on the data-generating process, which aligns with the complex relationships between financial variables. We use RCIT as an efficient implementation for KCIT. In general, we perform a hypothesis test on the squared Frobenius norm of a conditional cross-covariance operator $\Sigma_{\ddot{X}Y \cdot Z}$, where $\ddot{X} = (X, Z)$. If the norm is equal to zero, then *X* and *Y* are conditionally independent given the set *Z*. The output from RCIT is a p-value. If the p-values for multiple conditional independence tests are uniformly distributed between 0 and 1, we accept the null hypothesis that *X* and *Y* are conditionally or unconditionally independent. Otherwise, if p-values concentrate around zero, we reject the null hypothesis.

We run the conditional independence tests by cross-sections only (i.e. month by month). Since the distribution of p-values is not linear, direct averaging may produce misleading results. Therefore, we use the $\mathcal{X}^2$-test to assess the uniformity of the p-values distribution to accept or reject a hypothesis. To increase the power of the test, each cross-section is split into two subsamples (each contains approximately 1000 samples) to increase the sample size.

We present the uniformity test results for the p-value distribution in conditional and unconditional independence tests. To clearly explain the conditional independence test analysis, we split it into two cases: (1). If there exists a causal path from *X* to *Y* on the causal graph, and there is no conditional independence between *X* and *Y* that we can test due to graph structure or unobserved variables, we validate its existence by removing the causal path (counterfactual approach). If the p-value of the uniformity test is greater than 0.1 (uniform) for the conditional independence tests $X \perp\!\!\!\perp Y \mid Z$, and smaller than 0.1 (non-uniform) for the unconditional independence tests $X \perp\!\!\!\perp Y$, then the dependence between the two variables is removed by the conditioning set *Z*. The association between *X* and *Y* is considered spurious and we need to reassess the assumptions on the causal relationships in our SCM (the link between *X* and *Y* is not supported by data). (2). If there exists a causal path from *X* to *Y* and we have proper conditional independence conditions to test on, the same test results as in (1) indicate that empirical data support the causal path (since we are not testing the alternative hypothesis without the link).

In the following sections, we validate our SCM through conditional independence tests on each determinant (firm characteristics) of carbon emissions and each determinant of stock returns, and finally focus on the association between emissions and stock returns.

4.2.1 Determinants of Emissions

In our causal structure hypothesis (Figure 4.2), we assume sales, investment, and tangible assets to be the direct cause of emissions. Book-to-market ratio, size, and leverage also indirectly influence emissions. We justify the associations between these variables and emissions using conditional independence tests.

4.2.1.1 Leverage

Leverage can indirectly drive emissions through its impact on firm investment and hence sales. By providing the necessary capital for production expansion, leverage often results in increased emissions due to higher production levels. From our causal structure hypothesis (Figure 4.2), there is no direct causal path from leverage to emissions. Does the data support such an assumption? By applying the d-separation, the graph without a link between emissions and leverage entails the following conditional independence[4]: *Emissions* $\perp\!\!\!\perp$ *LEVERAGE* | *INVEST/A*, $SIZE_{t_0}$, PPE_{t_0}. As shown in Table 4.1, data does not support the causal graph without a link between emissions and leverage. To explain the association between leverage and emissions, we propose that the pressure to generate short-term revenue to meet debt obligations can lead firms to focus on maximizing output and put less attention on reducing environmental impact. We modify our SCM by adding the link *LEVERAGE* $\rightarrow$ *Emissions*. The updated SCM is presented in Figure 4.4.

4.2.1.2 Size

In our causal model presented in Figure 4.2, size is assumed to indirectly influence emissions through its impact on sales and investment-to-asset ratio. To check that there is no direct influence from size to emissions, we test the conditional independence entailed in Figure 4.2: *Emissions* $\perp\!\!\!\perp$ $SIZE_{t_0}$ | *INVEST/A*, *LEVERAGE*, *SALES*, PPE_{t_0}. We present the histogram for the p-values of the cross-sectional conditional independence tests in Figure 4.3 panel (a). P-values for both conditional and unconditional independence tests concentrate around zero. This implies that size and emissions are likely dependent and this dependence is not removed by the conditioning set. Therefore, the assumption that size only influences emissions indirectly through sales and investment-to-asset ratio is not supported by data. Based on the test results, we should consider adding a direct link from size to emissions in our causal model. The revised SCM is presented in Figure 4.4. Although larger firms may benefit from economies of scale, greater firm size may directly result in higher emissions due to complex supply chains and resource utilization.

4.2.1.3 Sales

Sales is an index that directly indicates the production scale of a company, we assume it to be the direct cause of firm emissions in our causal structure hypothesis (Figure 4.4). From the graph, the only path from sales to emissions is the direct link *SALES* $\rightarrow$ *Emissions*. By applying d-separation, there is no node that can block this direct path, so there is no conditional independence condition between sales and emissions. To resolve this issue, we take the counterfactual approach and remove the link *SALES* $\rightarrow$ *Emissions*. The causal graph in Figure 4.2 without the link from sales to emissions satisfies the following conditional independence: *Emissions* $\perp\!\!\!\perp$ *SALES* | *INVEST/A*, $SIZE_{t_0}$ The test results in Table 4.1 show that emissions and sales are not independent after conditioning. Thus, we reject the SCM without the link between sales and emissions. The data support the existence of *SALES* $\rightarrow$ *Emissions*.

4.2.1.4 Investment-to-asset Ratio

Similar to sales, we assume the investment-to-asset ratio directly influences emissions, as we believe firms may invest in green technology or improvements in efficiency, which impact carbon

[4]While d-separation gives us a set of conditioning sets making two variables conditionally independent, it is sufficient for us to conduct conditional independence test on the minimal necessary conditioning set.

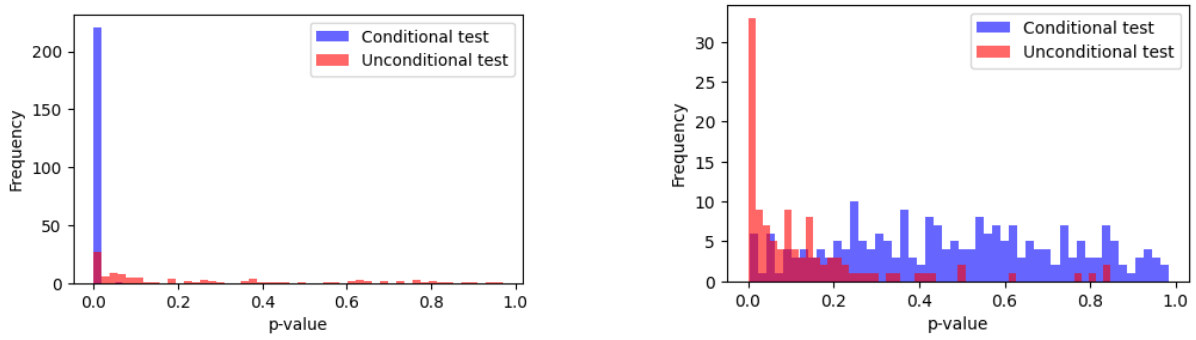
emissions. However, the existence of the link *INVEST/A* $\rightarrow$ *Emissions* prevents us from obtaining conditional independence conditions from the causal graph. We remove this link and test the alternative hypothesis which should satisfy *Emissions* $\perp\!\!\!\perp$ *INVEST/A* | *SALES*, *LEVERAGE*, $SIZE_{t_0}$, PPE_{t_0}. Test result in Table 4.1 indicates that the p-values of the cross-sectional conditional independence tests concentrate around zero, leading us to reject the alternative hypothesis.

4.2.1.5 Property, Plant, and Equipment

Firms with more property, plant, and equipment (PPE) tend to generate more carbon emissions due to the operation of machinery and facilities. Thus, we assume PPE is a direct cause of emissions. If the link from PPE to emissions does not exist, the following conditional independence must hold: *Emissions* $\perp\!\!\!\perp$ PPE_{t_0} | *INVEST/A*, *LEVERAGE*, $SIZE_{t_0}$. The test result in Table 4.1 supports the existence of the link PPE_{t_0} $\rightarrow$ *Emissions*.

4.2.1.6 Book-to-market Ratio

The book-to-market ratio is assumed to have indirect effects on emissions through its influence on a firm's investment strategies and leverage. High B/M ratio firms (undervalued firms) often adopt aggressive investment strategies to exploit growth opportunities. At the same time, they may find it more challenging to raise equity and might rely more on debt financing to fund their investments. If the direct link B/M_{t_0} $\rightarrow$ *Emissions* does not exist, the following conditional independence holds true: B/M_{t_0} $\perp\!\!\!\perp$ *Emissions* | *INVEST/A*, *LEVERAGE*, PPE_{t_0}, $SIZE_{t_0}$. As shown in Table 4.1, the p-value for the uniformity test is 0.19 $\geq$ 0.1. Therefore, we conclude that the data supports our causal graph in Figure 4.4 without a direct link between B/M and emissions. We present the histogram for p-values from cross-sectional conditional independence tests in Figure 4.3 panel (b).



(a) Size **(b)** B/M

Figure 4.3 Histograms of the p values of conditional and unconditional independence tests for size in panel (a) and B/M in panel (b). In panel (a), we test *Emissions* $\perp$ $SIZE_{t_0}$ | *INVEST/A*, *LEVERAGE*, *SALES*, PPE_{t_0} and *Emissions* $\perp$ $SIZE_{t_0}$. In panel (b), we test B/M_{t_0} $\perp\!\!\!\perp$ *Emissions* | *INVEST/A*, *LEVERAGE*, PPE_{t_0}, $SIZE_{t_0}$ and B/M_{t_0} $\perp\!\!\!\perp$ *Emissions*.

4.2.1.7 Other firm characteristics

Stock return volatility, return on equity, and earnings per share growth rate are not determinants of emissions in our causal structure hypothesis (Figure 4.4). To further verify the associations between these variables and emissions, we test the following conditional independencies: (1).

Emissions ⫫ *VOLAT* | *LEVERAGE* (2). *Emissions* ⫫ *ROE* | $LOGSIZE_{t_0}$, *SALES* (3). *EPSGR* ⫫ *Emissions* | $LOGSIZE_{t_0}$, *SALES* (4). *EPSGR* ⫫ *Emissions* | *ROE*. The test results are shown in Table 4.1. For volatility, p-values for both unconditional and conditional independence tests tend to be uniformly distributed between zero and one. These uniformly distributed p-values indicate that there is no significant association between volatility and emissions. For ROE and EPSGR, p-values for unconditional independence tests (*Emissions* ⫫ *ROE* and *Emissions* ⫫ *EPSGR*) concentrate around zero, while the p-values for the conditional independence tests are uniformly distributed. This pattern indicates that, when not conditioned on other variables, there appears to be some association between ROE, EPSGR, and emissions. However, this dependence disappears when we condition on relevant variables. These results suggest that the apparent associations observed in the unconditional tests are due to confounding factors. Therefore, we conclude that the data support the absence of obvious contemporaneous associations between these firm characteristics and emissions

Hypothesis	χ^2
Emissions ⊥ $SIZE_{t_0}$ | *INVEST/A*, *LEVERAGE*, *SALES*, PPE_{t_0}	0.0 → 0.00
Emissions ⫫ *LEVERAGE* | *INVEST/A*, $SIZE_{t_0}$, PPE_{t_0}	0.0 → 0.00
Emissions ⫫ *SALES* | *INVEST/A*, $SIZE_{t_0}$	0.0 → 0.00
Emissions ⫫ *INVEST/A* | *SALES*, *LEVERAGE*, $SIZE_{t_0}$, PPE_{t_0}	0.0 → 0.00
Emissions ⫫ PPE_{t_0} | *INVEST/A*, *LEVERAGE*, $SIZE_{t_0}$	0.0 → 0.00
B/M_{t_0} ⫫ *Emissions* | *INVEST/A*, *LEVERAGE*, PPE_{t_0}, $SIZE_{t_0}$	0.0 → 0.19
Emissions ⫫ *VOLAT* | *LEVERAGE*	0.11 → 0.26
Emissions ⫫ *ROE* | $LOGSIZE_{t_0}$, *SALES*	0.0 → 0.31
EPSGR ⫫ *Emissions* | $LOGSIZE_{t_0}$, *SALES*	0.0 → 0.24
EPSGR ⫫ *Emissions* | *ROE*	0.03 → 0.27

Table 4.1 Conditional independence test results for determinants of emissions[5]. The distribution of p-values for each conditional independence test can be found in Appendix B. For each test, we present the p-value from $\mathcal{X}^2$-test to assess uniformity in p-values. P-values for uniformity are presented as unconditional → conditional in the second column. If $p \geq 0.1$, we accept uniformity in p-values distribution from cross-sectional conditional independence tests. This uniformity indicates conditional independence between variables we test on.

In conclusion, we validate the relationships between firm characteristics and emissions in the SCM proposed in section 4.1. Through conditional independence tests, we find that data supports size and leverage to also be the direct causes of emissions, which are not included in our initial SCM (Figure 4.2). Based on these findings, we revise our causal model to include direct links from size and leverage to emissions, as illustrated in the updated SCM (Figure 4.4). These adjustments help our model to reflect the empirical data-generating process more accurately.

4.2.2 Determinants of Stock Returns

In this section, we perform conditional independence tests on determinants of stock return. We will leave the tests for emissions (which is also assumed to be one of the determinants) to the sext subsection as it is the main variable of interest.

In our SCM, we assume all firm characteristics prior to time t_1 ($SIZE_{t_0}$, PPE_{t_0}, B/M_{t_0}, *INVEST/A*, *LEVERAGE*, *SALES*, *ROE*, *EPSGR*, *Emissions*) would influence investor sentiment. Investor sentiment then impacts future stock returns (RET_{t_1}) either directly or indirectly through stock return volatility.

[5]We run each conditional independence test with cross-sectional emissions scope 1 and scope 2 data. The test results are similar for emissions in different scopes. Here we present the results of using scope 1 emission data.

We explain why each firm characteristic is assumed to be a driver of investor sentiment. Larger companies are often seen as more stable and competitive. They typically have a larger market share and more resources, which can boost investor confidence. Larger firms also tend to have diversified businesses and revenue streams, making them more resilient to market fluctuations and reducing investment risk. A high level of PPE indicates that a company has significant physical assets, often viewed as a sign of long-term growth potential. Companies with a high B/M ratio are seen as undervalued and having significant growth potential, leading investors to believe in their potential for value appreciation. A higher INVEST/A ratio signals active investment in growth and innovation, which usually positively influences investor sentiment. On the other hand, high leverage can indicate aggressive growth strategies, attracting investors seeking high returns, but it can also raise concerns about financial risk. Strong sales figures demonstrate robust revenue generation, enhancing investor confidence in the firm's profitability and market position. High ROE reflects that a company is effectively using shareholders' equity to generate profits, which are highly attractive to investors. Rapid EPS growth indicates increasing profitability, generating optimism among investors about future performance. Finally, if investors are concerned with the sustainability and social responsibility of a firm, a change in emissions will impact investor perception of the firm's performance.

It is important to note that investor sentiment is an unobserved variable, hence we can not test causal links connected to it. Therefore, we take the counterfactual approach to conduct conditional independence tests.

We take size as an example. Size is indirectly associated with stock returns through the unobserved mediator investor sentiment. Applying d-separation, there is no node we can condition on to block the path $SIZE_{t_0} \rightarrow$ *Investor Sentiment* $\rightarrow RET_{t_1}$ (we cannot condition on investor sentiment as it is unobserved). Thus, we can find no conditional independence condition between size and stock returns because we cannot make the two variables d-separated with the presence of the unobserved link $SIZE_{t_0} \rightarrow$ *Investor Sentiment*. To conduct the conditional independence test, we remove this link and test the alternative hypothesis (Figure 4.4 without the link $SIZE_{t_0} \rightarrow$ *Investor Sentiment*. If the alternative hypothesis true, the following conditional independence must hold: $SIZE_{t_0} \perp\!\!\!\perp RET_{t_1} \mid B/M_{t_0}$, *Emissions*, *INVEST/A*, *LEVERAGE*, PPE_{t_0}, *ROE*, *SALES*. The test result in Table 4.2 shows that the data does not support the alternative hypothesis. Similarly, to validate the associations between other firm characteristics (PPE_{t_0}, B/M_{t_0}, *SALES INVEST/A*, *VOLAT*, *ROE*, *EPSGR*) and stock returns, we take a similar approach. By removing the link from each firm characteristic, we obtain a set of conditional independencies. The test results are presented in Table 4.2. The data supports those firm-specific variables to be determinants of stock returns.

Note that $VOLAT_{t_1}$ is also assumed to be a driver of stock returns as increased volatility typically reflects higher risk, which can lead to greater uncertainty and potential for both higher returns and losses. This is supported by several frameworks including the Capital Asset Pricing Model (CAPM), risk-return tradeoff principle, and behavioral finance (Sharpe 1964, Fama & MacBeth 1973, Barberis & Thaler 2003). $VOLAT_{t_1}$ is directly connected to RET_{t_1}, so we need to remove this link to create testable conditional independence. However, there is another path between contemporaneous volatility and stock returns: $VOLAT_{t_1} \leftarrow$ *Investor Sentiment* $\rightarrow RET_{t_1}$. To block this path (which is a fork structure), we need to include the middle node investor sentiment in the conditioning set, whereas investor sentiment is an unobserved variable. Thus, to obtain the conditional independence condition between volatility and stock returns, we need to remove both $VOLAT_{t_1} \rightarrow RET_{t_1}$ and *Investor Sentiment* $\rightarrow VOLAT_{t_1}$. If these two links do not exist, the following conditional independence holds: $RET_{t_1} \perp\!\!\!\perp VOLAT_{t_1} \mid$ *LEVERAGE*. The test result in Table 4.2 shows that both the conditional and unconditional independent tests p-values concentrate around zero. Contemporaneous volatility and stock returns are unconditionally dependent, and their dependence is not removed by leverage. This supports that there should

be a causal path between them, but we cannot tell from tests that the existing path is $VOLAT_{t_1} \rightarrow RET_{t_1}$, $VOLAT_{t_1} \leftarrow$ *Investor Sentiment* $\rightarrow RET_{t_1}$, or both of them.

Another variable that we encounter difficulty in testing conditional independence is leverage. By applying d-separation, we notice two conflicting paths considering the conditioning set that d-separate leverage and stock returns: (1). $LEVERAGE \rightarrow VOLAT_{t_1} \rightarrow RET_{t_1}$ (2). $LEVERAGE \rightarrow VOLAT_{t_1} \rightarrow$ *Investor Sentiment* $\rightarrow RET_{t_1}$. In the first path, $VOLAT_{t_1}$ is the middle node of a chain, so we need to condition on it to block this path. In the second path, $VOLAT_{t_1}$ becomes a middle node of a collider (note that we cannot condition on investor sentiment to block this path, as it is unobserved). Thus, we cannot obtain any testable conditional independence between leverage and stock returns.

Removing the node $VOLAT_{t_1}$ will resolve this issue. If $VOLAT_{t_1}$ is removed, we can test the association between leverage and stock returns in a similar counterfactual approach as we do with size. By removing the unobserved link *LEVERAGE* $\rightarrow$ *Investor Sentiment*, the following condition should hold: $LEVERAGE \perp\!\!\!\perp RET_{t_1} \mid B/M_{t_0}$, *Emissions*, *INVEST/A*, $LOGPPE_{t_0}$, $LOGSIZE_{t_0}$, *SALES*, *VOLAT*. The test result in Table 4.2 tells that data support leverage to be a determinant of stock returns.

We present the SCM with $VOLAT_{t_1}$ being removed in Figure 4.5. Despite volatility at t_1 adding interpretability to our SCM as it is an important indicator reflecting investor sentiment, we cannot fully justify its role in our model. With the model presented in Figure 4.5, the association between each firm characteristics (except emissions) and stock returns are properly validated through conditional independence tests and are supported by data.

Hypothesis	χ^2
$SIZE_{t_0} \perp\!\!\!\perp RET_{t_1} \mid B/M_{t_0}$, *Emissions*, *INVEST/A*, *LEVERAGE*, PPE_{t_0}, *ROE*, *SALES*	$0.0 \rightarrow 0.01$
$PPE_{t_0} \perp\!\!\!\perp RET_{t_1} \mid B/M_{t_0}$, *Emissions*, *INVEST/A*, *LEVERAGE*, $SIZE_{t_0}$, *SALES*	$0.0 \rightarrow 0.00$
$B/M_{t_0} \perp\!\!\!\perp RET_{t_1} \mid$ *INVEST/A*, *LEVERAGE*, PPE_{t_0}, $SIZE_{t_0}$	$0.0 \rightarrow 0.00$
$RET_{t_1} \perp\!\!\!\perp$ *SALES* $\mid$ *Emissions*, *INVEST/A*, *LEVERAGE*, PPE_{t_0}, $SIZE_{t_0}$, *ROE*	$0.0 \rightarrow 0.00$
INVEST/A $\perp\!\!\!\perp RET_{t_1} \mid B/M_{t_0}$, *Emissions*, *LEVERAGE*, PPE_{t_0}, $SIZE_{t_0}$, *SALES*	$0.0 \rightarrow 0.00$
$RET_{t_1} \perp\!\!\!\perp$ *VOLAT* $\mid$ *LEVERAGE*	$0.0 \rightarrow 0.00$
$RET_{t_1} \perp\!\!\!\perp$ *ROE* $\mid$ *EPSGR*, $SIZE_{t_0}$, *SALES*	$0.0 \rightarrow 0.00$
EPSGR $\perp\!\!\!\perp RET_{t_1} \mid$ *ROE*	$0.0 \rightarrow 0.08$
$RET_{t_1} \perp\!\!\!\perp VOLAT_{t_1} \mid$ *LEVERAGE*	$0.0 \rightarrow 0.00$
LEVERAGE $\perp\!\!\!\perp RET_{t_1} \mid B/M_{t_0}$, *Emissions*, *INVEST/A*, PPE_{t_0}, $SIZE_{t_0}$, *SALES*, *VOLAT*	$0.0 \rightarrow 0.00$

Table 4.2 Conditional independence test results for determinants (except emissions) of stock returns. The distribution of p-values for each conditional independence test can be found in Appendix B. For each test, we present the p-value from $\mathcal{X}^2$-test to assess uniformity in p-values. P-values for uniformity are presented as unconditional $\rightarrow$ conditional in the second column. If $p \geq 0.1$, we accept uniformity in p-values distribution from cross-sectional conditional independence tests. This uniformity indicates conditional independence between variables we test on.

4.2.3 Emission Variables

In this section, we justify the role of emissions as a determinant of stock returns. The carbon emissions of a company are evaluated in three emission variables: total emissions, emission growth rate, and emission intensity. Following the study in Bolton & Kacperczyk (2021), we consider emissions in both scope 1 and scope 2.

Similar to the example of size (which is discussed in the previous section), emissions indirectly influence stock returns through unobserved investor sentiment. To obtain testable conditional independence between emissions and stock returns, we counterfactually remove the

link *Emissions* $\rightarrow$ *Investor Sentiment* (red dashed link in Figure 4.5). If this link does not exist, the following conditional independence must be satisfied by the data: *Emissions* $\perp\!\!\!\perp$ RET_{t_1} | *INVEST/A*, *LEVERAGE*, PPE_{t_0}, $SIZE_{t_0}$, *SALES*. We substitute emissions into total emissions scope 1, total emissions scope 2, emission growth rate scope 1, and emission growth rate scope 2, and test the association between each of these variables and stock returns. The test results are shown in Table 4.3. The results indicate that the link between emissions and investor sentiment exists for all four emission variables.

By validating the relationship between emissions and stock returns, we prove the existence of the causal link *Emissions* $\rightarrow$ *Investor Sentiment* (red dashed link in Figure 4.5). This answers the central question studied in Bolton & Kacperczyk (2021): *Do investors care about emissions?* The answer from our structural causal model is yes. We will estimate the magnitude of the associations between emission variables and returns in chapter 5.

Note that for carbon intensity, we cannot justify its association with stock returns. As explained in section 4.1, carbon intensity (ratio of emissions to sales) is totally determined by emissions and sales, and there is no stochastic exogenous variable that causes it. Hence emission intensity is not a rigorous endogenous variable in SCM and we include it in our causal model (Figure 4.5) just for illustrative purposes. Consequently, it is also not valid to perform conditional independence tests between emission intensity and stock returns. This is a limitation of our SCM and will be further discussed in chapter 6.

Hypothesis	χ^2
Total Scope1 $\perp\!\!\!\perp$ RET_{t_1} | *INVEST/A*, *LEVERAGE*, PPE_{t_0}, $SIZE_{t_0}$, *SALES*	0.0 $\rightarrow$ 0.00
Total Scope2 $\perp\!\!\!\perp$ RET_{t_1} | *INVEST/A*, *LEVERAGE*, PPE_{t_0}, $SIZE_{t_0}$, *SALES*	0.0 $\rightarrow$ 0.00
Growth Scope1 $\perp\!\!\!\perp$ RET_{t_1} | *INVEST/A*, *LEVERAGE*, PPE_{t_0}, $SIZE_{t_0}$, *SALES*	0.05 $\rightarrow$ 0.00
Growth Scope2 $\perp\!\!\!\perp$ RET_{t_1} | *INVEST/A*, *LEVERAGE*, PPE_{t_0}, $SIZE_{t_0}$, *SALES*	0.0 $\rightarrow$ 0.00

Table 4.3 Test results for conditional independence between emissions and stock returns. The distribution of p-values for each conditional independence test can be found in Appendix B. For each test, we present the p-value from $\mathcal{X}^2$-test to assess uniformity in p-values. P-values for uniformity are presented as unconditional $\rightarrow$ conditional in the second column. If p $\geq$ 0.1, we accept uniformity in p-values distribution from cross-sectional conditional independence tests. This uniformity indicates conditional independence between variables we test on.

4.3 Control Set Selection

In chapter 3, we reproduce the baseline regression in Bolton & Kacperczyk (2021) to investigate the association between emissions and stock returns. In section 4.1 and section 4.2, we develop and validate a structural causal model (Figure 4.5) for the causal relationships between emissions, stock returns, and other firm characteristics.

As discussed in section 2.4, using control variables based on untested beliefs about what explains the dependent variable can introduce biases into regression results. While adding more control variables generally increases the model's goodness of fit (R^2), it may lead to biased estimates for the coefficient of the variable of interest. For instance, although $VOLAT_{t_1}$ is removed in the final validated SCM, we can observe from Figure 4.4 that emissions drive stock returns through investor sentiment. Volatility at time t_1 is a descendent of the mediator *Investment Sentiment*. Thus, including volatility at time t_1 in the control set will lead to bias in estimating the total effect of emissions on stock return. Applications of the backdoor criterion help us avoid including those variables that cause biases in regression estimation results (see sections 2.2 and 2.3 for more detailed discussion).

The baseline regression in section 3.2 following the study Bolton & Kacperczyk (2021) does not justify its use of control variables, which can potentially induce biases in regression results. Based on the SCM we have developed, we apply the backdoor criterion to identify the minimal adjustment set we need to control for to estimate the unbiased total effect of emissions on stock returns. The justified control variables for stock returns in time t_1 are *INVEST/A*, *LEVERAGE*, PPE_{t_0}, $SIZE_{t_0}$, *SALES*[6]. This new control set blocks all spurious paths between emissions and stock returns and avoids collider bias and mediator bias in estimating the total causal effect. We can now turn to the general results by re-estimating the unbiased associations between emissions and stock returns using the reasoned control set.

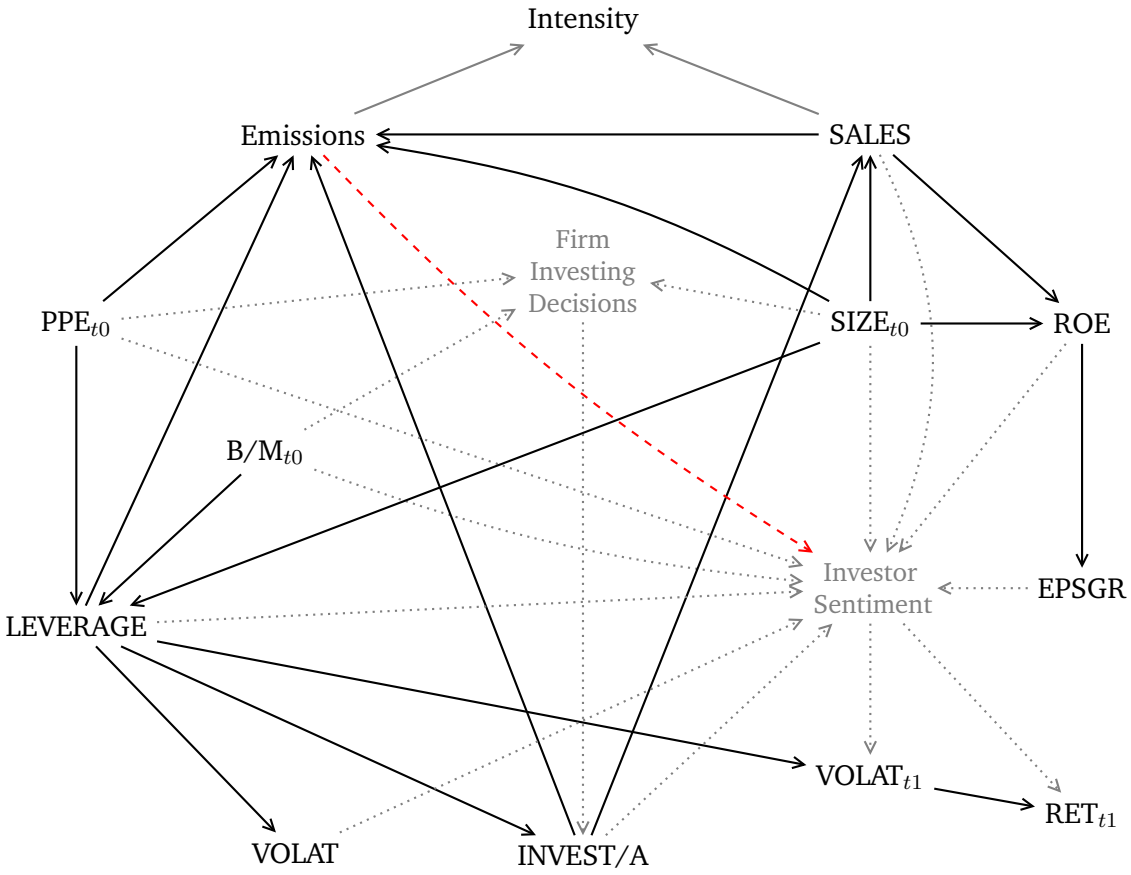


Figure 4.4 The updated SCM for carbon emissions and stock returns after conditional independence tests. Compared to Figure 4.2, we add two direct link to the graph: $SIZE_{t_0} \rightarrow$ *Emissions* and *LEVERAGE* $\rightarrow$ *Emissions*. Firm investing decisions and investor sentiment are two unobserved variables and are painted in gray. The directed edges connected to those two unobserved mediators are drawn as dotted gray lines. Time t_0 is the initial time, t_1 is the final time, and variables without a time mark are considered to fall within the interval between t_0 and t_1. The unobserved edge from *Emissions* to *Investor Sentiment* marked in red dashed line is the essential unobserved link we are interested in. Emission intensity is not a rigorous endogenous variable, but a derived quantity totally determined by sales and emissions. The edge linked to Intensity is marked differently in gray.

[6]Note that it is a coincide that the minimal adjustment set from the backdoor criterion is the same as the (minimal sufficient) conditioning set we identified through d-separation in Table 4.3. Conditioning sets identified from d-separation need to meet further criteria to be the minimal adjustment set. See section 2.2 for more details.

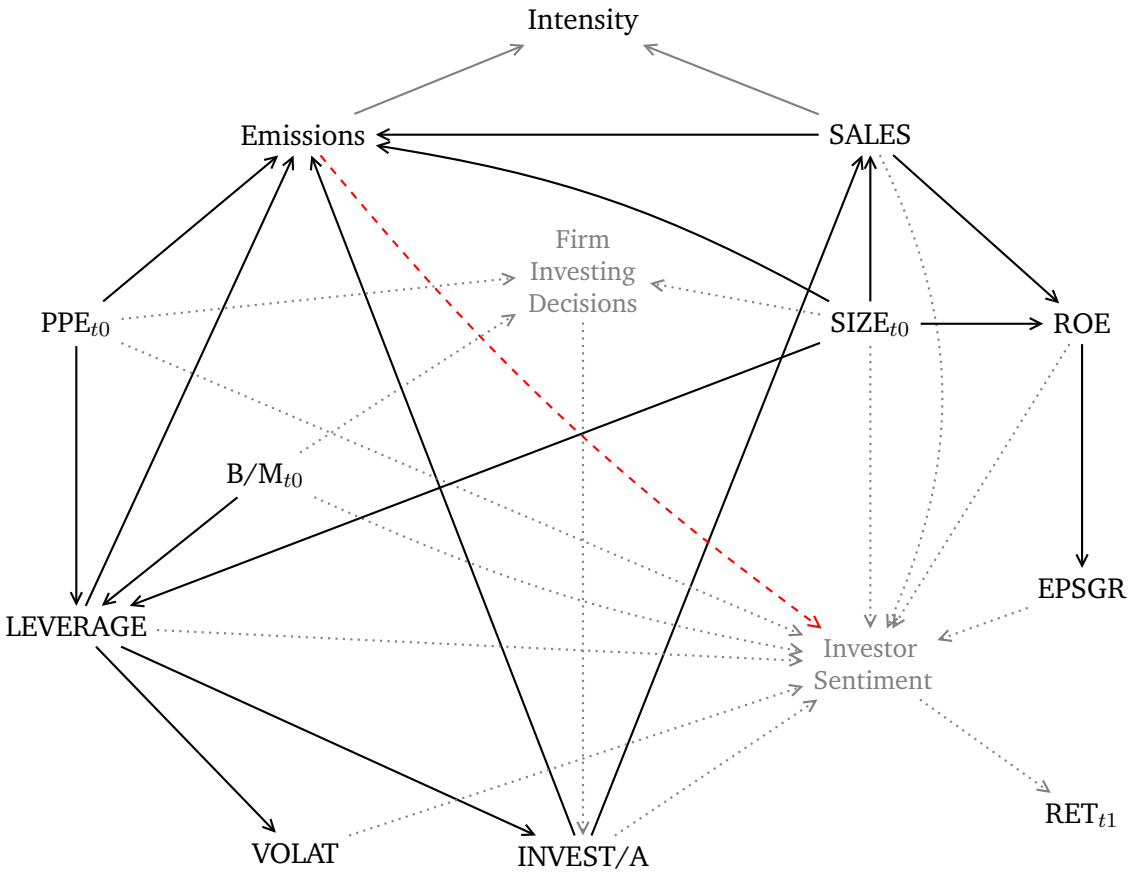


Figure 4.5 The final validated SCM for carbon emissions and stock returns. The node *VOLAT*$_{t_1}$ is removed from the graph as we cannot justify its role in the model through conditional independence tests. Firm investing decisions and investor sentiment are two unobserved variables and are painted in gray. The directed edges connected to those two unobserved mediators are drawn as dotted gray lines. Time t_0 is the initial time, t_1 is the final time, and variables without a time mark are considered to fall within the interval between t_0 and t_1. The unobserved edge from *Emissions* to *Investor Sentiment* marked in red dashed line is the essential unobserved link we are interested in. Emission intensity is not a rigorous endogenous variable, but a derived quantity totally determined by sales and emissions. The edge linked to Intensity is marked differently in gray.

4.4 Re-estimate Associations

Using the results from our SCM, we now re-estimate the unbiased total causal effect of emissions on stock returns. Compared to the baseline regression in section 3.2, we use a justified control set *Controls'*$_{i,t^*}$. Additionally, we adjust the emissions variable to be measured at the time prior to the measurement of stock returns. We propose that it is more meaningful to investigate the effect of past emissions on stock returns compared to their contemporaneous associations, as investors take time to receive and react to firm emissions information. This change aligns with the logic in our SCM as well as the central question to answer: *Do investors care about carbon emissions?* The new regression model is as follow:

$$\textit{RET}_{i,t+1} = a_0 + a_1 \textit{Emissions}_{i,t} + a_2 \textit{Controls}_{i,t^*} + \mu_t + \sigma_{industry} + \epsilon_{i,t} \tag{4.1a}$$

$$\textit{Controls'}_{i,t^*} = (\textit{INVEST/A}_{i,t}, \textit{LEVERAGE}_{i,t}, \textit{PPE}_{i,t-1}, \textit{SIZE}_{i,t-1}, \textit{SALESGR}_{i,t}) \tag{4.1b}$$

where the dependent variable *RET*$_{i,t+1}$ represents the monthly stock return for firm i in month $t+1$, and *Emissions*$_{i,t}$ refers to one of the emission variables: natural logarithm of total firm-level emissions and emissions growth rate in Scope 1 and Scope 2. The new control set *Controls'*$_{i,t^*}$ is defined in Eq.(4.1b), where t^* denotes the specific time period relevant for each control variable. *INVEST/A*$_{i,t}$ is the investment-to-asset ratio in month t, *LEVERAGE*$_{i,t}$ represents the leverage of firm i in month t, *PPE*$_{i,t-1}$ (property, plant, and equipment) and *SIZE*$_{i,t-1}$ are normalized using natural logarithm and are measured in month $t-1$, providing a view of the firm's physical capital over a longer period. *SALESGR*$_{i,t}$[7] is the sales growth figure in month t. Finally, μ_t represents year-month fixed effects, $\sigma_{industry}$ denotes industry fixed effects, and $\epsilon_{i,t}$ is the error term. The results for the re-estimated associations between carbon emissions and stock returns are presented in the next chapter.

[7] We use sales growth rate to be a measurement for *SALES* in the SCM because it captures the growth aspect of sales, which is more dynamic and indicative of a firm's performance changes over time. Additionally, *SALESGR* was used in the baseline regression, providing consistency with previous analyses and ensuring comparability of results.

Chapter 5

Results

In Table 5.1, we present the results of the new regression model defined in Eq.(4.1) with the justified control set *Controls'*$_{i,t^*}$ (specified in Eq. 4.1b). Note that by controlling for the minimal adjustment set between emissions and stock returns according to our SCM developed in chapter 4, we estimate the total causal effect of emissions (measured as total emissions or emission growth rate) on stock return.

Panel A shows the re-estimated coefficients for the natural logarithm of total emissions in Scope 1 and 2 from the new regression model. Scope 1 emissions have no significant association with stock returns in all specifications, while Scope 2 emissions consistently show a significant negative effect on stock returns across all specifications. For emission growth rates in panel B, Scope 1 emission growth rates have a significant positive relationship with stock returns across all columns, though the significance level varies. In contrast, Scope 2 emission growth rates do not show any significant relationship with stock returns.

Compared with the results for the baseline regression (Table 3.2), Scope 1 total emissions showed a significant positive relationship in some specifications, while the new regression shows Scope 1 emissions have no significant effect on stock returns. Scope 2 total emissions in both the baseline and the new regression demonstrate a consistently significant negative relationship with stock returns. The statistical significance for this association is slightly less in the new regression if we look at the p-values for the coefficients.

In previous results, Scope 1 emission growth rate had no significant association (and the association was negative) with stock returns across all specifications. Oppositely, the results of the new regression suggest a significant positive relationship between Scope 1 emission growth and stock returns. Scope 2 emission growth rate shows no significant relationship with stock returns in both regressions.

From the baseline regression results, we inferred that emission levels might be more relevant to stock returns than emission growth rates due to the lack of significant associations observed for the latter. However, this inference is contradicted by the new regression results, which show that Scope 1 emission growth rates have a significant positive impact on stock returns.

To conclude, when accounting for estimates of total causal effects, the new regression model confirms the consistent negative impact of Scope 2 total emissions on stock returns while revealing new insights about the positive effect of Scope 1 emission growth rates.

Table 5.1 Re-estimated associations between stock returns and carbon emissions

The table shows the OLS coefficients for emission variables in the linear regression model mentioned in Eq. 4.1. We report the statistical significance p values in the parentheses below the coefficients. The

dependent variable is RET. The emission variable takes one of the following forms: log total emission and emission growth rate for both Scope 1 and Scope 2. Columns (1) and (2) show the regression coefficients without considering year-month or industry fixed effects, columns (3) and (4) additionally include the year-month effects, and columns (5) and (6) take all fixed effects into account. Panel A reports the results for the natural logarithm of total emissions; Panel B reports the results for growth rate in emissions. ***, **, * indicates statistical significance at 1%, 5%, and 10%, respectively.

Panel A: Total Emissions						
Variables	(1) RET	(2) RET	(3) RET	(4) RET	(5) RET	(6) RET
Log (Scope 1)	0.0244		0.0210		-0.0138	
	(0.339)		(0.393)		(0.662)	
Log (Scope 2)		-0.0930***		-0.0985***		-0.1138***
		(0.008)		(0.005)		(0.005)
Observations	80201	80201	80201	80201	80201	80201
Year-month F. E.	No	No	Yes	Yes	Yes	Yes
Industry F. E.	No	No	No	No	Yes	Yes
Panel B: Emission Growth Rates						
Growth Scope1	1.9240***		1.1090*		1.4046**	
	(0.002)		(0.058)		(0.016)	
Growth Scope2		0.7421		0.3914		0.4808
		(0.229)		(0.512)		(0.418)
Observations	80201	80201	80201	80201	80201	80201
Year-month F. E.	No	No	Yes	Yes	Yes	Yes
Industry F. E.	No	No	No	No	Yes	Yes

Chapter 6

Discussion

In this study, we explored the relationship between emissions and stock returns which is a widely studied topic in empirical finance research. While many studies (Bolton & Kacperczyk 2021, Aswani et al. 2024) adopt a purely statistical approach to evaluate this relationship, we tackle this question using a systematical causal framework.

By developing a causal model for the emission, stock returns, and other relevant firm characteristics, we clearly stated our hypotheses on the causal relationships between variables. We then explicitly validated those relationships by conditional independence tests. In our SCM (Figure 4.5), the role of emissions in determining future stock returns was justified by the validated causal path *Emissions* $\rightarrow$ *Investor Sentiment* $\rightarrow$ RET_{t_1}. Therefore, based on our causal model, the qualitative association between emissions and stock returns exists and is supported by data. Using the backdoor criterion, we select the minimal adjustment set between emissions and stock returns as the control set, which should avoid biases (eg. collider bias and mediator bias) in the estimated associations. Finally, we re-estimated the associations between emission variables and stock returns and compared the results with the baseline regression.

In contrast to regression, which makes no assumptions about underlying causality in data and simply identifies the statistical associations between variables, we are able to estimate the unbiased total causal effect and better explain the causal story between relevant variables through the SCM approach.

Our results revealed a significant positive relationship between Scope 1 emission growth rate and stock returns, which was negative and not significant in the baseline regression. Additionally, Scope 1 total emissions showed a less significant positive relationship with stock returns compared to the baseline estimates. These differences highlight the potential biases in the baseline estimates.

Compared to the baseline regression following Bolton & Kacperczyk (2021), our new regression model with a systematically justified control set is more transparent. The SCM approach allows us to avoid spurious associations and provides clearer insights into the causal relationships between emissions and stock returns. We now briefly discuss a number of important limitations in our derived SCM.

6.1 Limitations of SCM

A notable limitation of our study is related to carbon intensity. We cannot properly put it as an endogenous variable in our SCM as it is totally determined by other variables in the graph and this violates the definition for endogenous variables. As a consequence, our causal model is not

able to specify a reasoned control set for emission intensity, and thus we cannot re-estimate the association between emission intensity and stock returns in our new regression model.

Besides, we assume all firm characteristics in SCM are in between the time interval $[t_0, t_1]$. In reality, firm investing decisions and investor sentiment do not occur in a vacuum and are influenced by a myriad of factors that unfold over time. For example, strategic decisions made by firms regarding investment in production and emissions policies often span multiple periods and are influenced by expectations about future market conditions. Similarly, market reactions to firm-level information can be immediate or delayed, influenced by information diffusion processes, investor sentiment, and macroeconomic conditions.

While we can only include limited variables in a causal graph, firms and investors can take much more variables into account when making investing decisions. In general, our model complexity cannot match the real-world complexity. This is a common issue in empirical finance studies. When approaching an empirical question through structural causal modeling, we focus on explaining the causal relationships in the model. As discussed above, this helps us to determine the actual magnitudes of the total causal effects.

Chapter 7

Conclusion

Understanding the causal structure between variables is crucial when estimating variable effects in empirical finance research. When we perform regressions to estimate the associations between two variables, using control variables without justification would not only reduce transparency in the model but also induce potential biases in the results.

In this report, we proposed a systematical approach based on structural causal modeling (SCM) to identify the control variables in a regression model. We developed a causal model for emissions and stock returns and explicitly validated the model using conditional independence tests. When performing conditional independence tests on SCM, we counterfactual prove the qualitative relationship between emissions and investor sentiment in our SCM is supported by data. Thus, we argue that investors do care about carbon emissions when they are investing in stocks. We then estimate the effect of emissions on stock returns from a causal perspective and compare the results to the baseline regression without a causal justification. Overall, we have shown that there are associations between emissions and stock return, but the significance of associations vary between different emission scopes and emission variables.

While involving more control variables in regression models can usually improve the statistical performance (eg. goodness of fit) of the model, this approach may lead to biased results. Furthermore, due to the lack of transparency in the model, it becomes challenging to identify these biases and interpret the results accurately. In this context, we have shown how to use SCM to resolve these issues. By laying down validated assumptions on the causality between variables and hence the underlying data-generating process, we can explicitly explain the story behind our model, allowing for a clearer and more accurate interpretation of the results.

Appendix A

First Appendix

A.1 Graph

A mathematical graph is a collection of *vertices* (or *nodes*) and *edges* (Pearl et al. 2016). The nodes are connected (or not) by the edges. Figure A.1 gives a simple example of a graph, where X, Y, and Z are nodes, and A and B are the edges connecting the nodes.

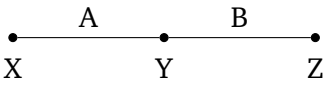


Figure A.1 An undirected graph

A *path* between two nodes is a sequence of nodes in which each node is connected by an edge to the next. In Figure A.1, there is a path from X to Z, because X is connected to Y, and Y is connected to Z.

Edges in a graph can be *directed* or *undirected*. Edges with a direction, indicated by an arrow are directed, and vice versa for undirected edges. A graph in which all of the edges are directed is a *directed graph*. The node that a directed edge starts from is called the *parent* of the node that the edge goes into; conversely, the node that the edge goes into is the *child* of the node it comes from (Pearl et al. 2016). In Figure A.2, X is the parent of Y, and Y is the parent of Z; similarly, Y is the child of X, and Z is the child of Y.

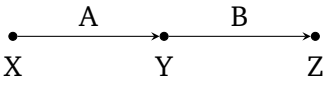


Figure A.2 A directed graph in which node X is a parent of Y and Y is a parent of Z

A path between two nodes is a *directed path* if it can be traced along the arrows. If two nodes are connected by a directed path, then the first node is the *ancestor* of every node on the path, and every node on the path is the *descendant* of the first node. For instance, in Figure A.2, X is the ancestor of both Y and Z, and both Y and Z are descendants of X (Pearl et al. 2016).

If a directed path exists from a node to itself, the path is called *cyclic*. A directed graph with no cycles is *acyclic*. Figure A.2 (a) is an acyclic graph since there is no path from any node to itself, whereas in Figure A.2 (b) there are directed paths from X to X, for example, so it is a cyclic graph.

Figure A.3 (a)acyclic graph; (b) cyclic graph

Appendix B

Second Appendix

All code for this project can be found at https://github.com/bslrjlb/M4R-code.git.

Bibliography

Aswani, J., Raghunandan, A. & Rajgopal, S. (2024), 'Are carbon emissions associated with stock returns?', *Review of Finance* **28**(1), 75–106.

Barberis, N. & Thaler, R. (2003), 'A survey of behavioral finance', *Handbook of the Economics of Finance* **1**, 1053–1128.

Bolton, P. & Kacperczyk, M. (2021), 'Do investors care about carbon risk?', *Journal of financial economics* **142**(2), 517–549.

Busch, T., Johnson, M., Pioch, T. & Kopp, M. (2018), 'Consistency of corporate carbon emission data', *Hamburg. Retrieved September* **29**, 2018.

Carhart, M. M. (1997), 'On persistence in mutual fund performance', *The Journal of finance* **52**(1), 57–82.

Cenci, S. & Kealhofer, S. (2022), 'A causal approach to test empirical capital structure regularities', *The Journal of Finance and Data Science* **8**, 214–232.

Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K. & Schölkopf, B. (2012), 'Inferring deterministic causal relations', *arXiv preprint arXiv:1203.3475* .

Fama, E. F. & French, K. R. (1995), 'Size and book-to-market factors in earnings and returns', *The journal of finance* **50**(1), 131–155.

Fama, E. F. & MacBeth, J. D. (1973), 'Risk, return, and equilibrium: Empirical tests', *Journal of political economy* **81**(3), 607–636.

Flesch, I. & Lucas, P. J. (2007), 'Markov equivalence in bayesian networks', *Advances in probabilistic graphical models* pp. 3–38.

Fukumizu, K., Bach, F. R. & Jordan, M. I. (2004), 'Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces', *Journal of Machine Learning Research* **5**(Jan), 73–99.

Fukumizu, K., Gretton, A., Sun, X. & Schölkopf, B. (2007), 'Kernel measures of conditional dependence', *Advances in neural information processing systems* **20**.

Greenland, S., Pearl, J. & Robins, J. M. (1999), 'Causal diagrams for epidemiologic research', *Epidemiology* **10**(1), 37–48.

Heckman, J. J. (2008), 'Econometric causality', *International statistical review* **76**(1), 1–27.

IBM team (2023), 'What are scope 3 emissions? — IBM, Think', https://www.ibm.com/topics/scope-3-emissions. [Online; accessed 30-June-2023].

In, S. Y., Park, K. Y. & Monk, A. (2017), 'Is "being green" rewarded in the market? an empirical investigation of decarbonization risk and stock returns', *International Association for Energy Economics (Singapore Issue)* **46**(48), 46–48.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B. & Schölkopf, B. (2012), 'Information-geometric approach to inferring causal directions', *Artificial Intelligence* **182**, 1–31.

Pearl, J. (1995), From bayesian networks to causal networks, *in* 'Mathematical models for handling partial knowledge in artificial intelligence', Springer, pp. 157–182.

Pearl, J. (2009), *Causality*, Cambridge university press.

Pearl, J., Glymour, M. & Jewell, N. P. (2016), *Causal inference in statistics: A primer*, John Wiley & Sons.

Penman, S. H. (2013), *Financial statement analysis and security valuation*, McGraw-hill.

Rajan, R. & Zingales, L. (1996), 'Financial dependence and growth'.

Runge, J. (2018), 'Causal network reconstruction from time series: From theoretical assumptions to practical estimation'.

Sharpe, W. F. (1964), 'Capital asset prices: A theory of market equilibrium under conditions of risk', *The journal of finance* **19**(3), 425–442.

Strobl, E. V., Zhang, K. & Visweswaran, S. (2019), 'Approximate kernel-based conditional independence tests for fast non-parametric causal discovery', *Journal of Causal Inference* **7**(1), 20180017.

Taagepera, R. (2008), *Making social sciences more scientific: The need for predictive models*, Oxford University Press, USA.

Titman, S. & Wessels, R. (1988), 'The determinants of capital structure choice', *The Journal of finance* **43**(1), 1–19.

Zhang, K., Peters, J., Janzing, D. & Schölkopf, B. (2012), 'Kernel-based conditional independence test and application in causal discovery', *arXiv preprint arXiv:1202.3775* .

Zhang, L. (2005), 'The value premium', *The Journal of Finance* **60**(1), 67–103.