Green Silence: Double machine learning carbon emissions under sample selection bias*

Cathy Yi-Hsuan Chen† Abraham Lioui‡ Olivier Scaillet§

January 11, 2026

Abstract

Voluntary carbon disclosure creates a paradox where firms strategically choose whether to report emissions, with high emitters often remaining silent to exploit estimation biases. This self-selection problem distorts climate risk pricing and policy analysis. We propose a three-step Heckman-inspired framework for high-dimensional data using kernel group lasso, double machine learning, and peer firm information. An empirical study of 3,444 U.S. firms (2010–2023) finds strong evidence of selection bias. Naïve imputation methods underestimate emissions, leading to a $2.6 billion carbon tax shortfall and up to $525 billion in hidden social costs. Correcting for non-disclosure bias reveals a higher carbon premium.

Keywords: carbon emissions, machine learning, sample selection.
JEL Classification: C12, C13, C33, C51, C52, C82, G11, G18, Q52, Q54, Q56, Q58.

*We thank C. Gaillac, W. Haerdle, S. Pouget, Z. Sautner, and participants to FinEML 2025, 2025 AIDA Workshop, EC^2 2025 Conference, Sustainable Finance Research Forum (Paris 2025), and to seminars at University of Zürich and University of Orléans for their insightful comments and suggestions.

†Adam Smith Business School, University of Glasgow, UK (*e-mail: CathyYi-Hsuan.Chen@glasgow.ac.uk*)
‡EDHEC Business School, France. (*e-mail: Abraham.Lioui@edhec.edu*)
§Université de Genève and Swiss Finance Institute, Switzerland. (*e-mail: Olivier.Scaillet@unige.ch*)

1 Introduction

Greenhouse gas emissions have reached new highs and climate impacts are intensifying globally, according to the UNEP Emissions Gap Report 2024.[1] Monitoring corporate carbon emissions is therefore a key ingredient for achieving Net Zero targets (Net Zero (2024)). Yet, when disclosure obligations are only beginning to emerge, firm-level emissions are largely observed through voluntary reporting and vendor-constructed datasets. Among data vendors, the Carbon Disclosure Project (CDP) is the most widely used by academics and practitioners and often serves as a backbone input for other providers. According to the CDP disclosure report, 23,188+ firms voluntarily disclosed climate-related information in 2023, a 140% increase relative to 2020.[2] Out of these firms, 8,000 (35%) disclosed for the first time. Yet, just under 400 companies (about 2%) were A listed by CDP, which is recognized for the very high quality of the published information.[3] While CDP coverage is voluntary, most emissions databases also rely heavily on vendor estimates for non-disclosing firms. Such pervasive reliance on imputed emissions implies that mismeasurement is not a marginal issue but a central feature of current carbon data. This concern aligns with a long-standing view in climate economics emphasizing that climate change is a global externality and one of the largest market failures precisely because risks, uncertainty, and non-linear damages tend to be understated when measurement is imperfect (Stern (2008)).

This paper studies a simple but consequential mechanism: when disclosure is voluntary and emissions for non-disclosers are imputed, the resulting carbon data are subject to endogenous sample selection. We introduce the term `green silence` as a reduced-form description of non-disclosure in voluntary carbon reporting regimes and show how it generates selection bias in observed emissions (Heckman (1979)). We then propose a Heckman-inspired three-step framework for high-dimensional settings that corrects for strategic non-disclosure and preserves variable selection consistency under selection. Methodologically, we integrate kernel group lasso (KG-lasso) - to learn carbon-relevant neighbourhoods from *carbon next door* - with double machine learning (DML) to obtain valid inference and robust testing in the presence of many covariates (Chernozhukov et al. (2018)). Empirically, using annual *Trucost* data on 3,444 U.S. firms from 2010 - 2023, we reject the null of no selection bias and document systematic underestimation of emissions when selection is ignored. Importantly, our results do *not* require a specific behavioral narrative (e.g., opportunism): underestimation arises mechanically whenever emissions are imputed using low-dimensional proxies and/or when the disclosure process is not modeled.

[1] https://www.unep.org/resources/emissions-gap-report-2024
[2] https://cdp.net/en/insights/cdp-2023-disclosure-data-factsheet.
[3] https://www.cdp.net/en/press-releases/scores-press-release-2023.

In the disclosure process, selection is natural. Firms' decisions to disclose reflect heterogeneous economic, strategic, and organizational considerations. Some firms may anticipate that third-party estimates misrepresent their true environmental footprint, while others may refrain from disclosure due to the absence of a clearly articulated transition path, concerns about future commitments, internal measurement constraints, or uncertainty about how disclosed information will be interpreted by investors, regulators, and other stakeholders. Green silence does not presume a unique motive or intent; it summarizes the aggregate outcome of these heterogeneous incentives. This endogenous participation implies that observed carbon footprints constitute a non-random sample of firms, giving rise to systematic selection bias in emissions data (Heckman (1979)). Just as Heckman observed that wage samples exclude workers whose reservation wages keep them out of the labor force, voluntary carbon datasets exclude firms whose emissions remain unobserved because they opt out of disclosure. Reported emissions data - like the earnings of migrants or trainees - therefore do not reflect the counterfactual: what non-disclosers would have reported under mandatory transparency. Conventional comparisons (disclosers vs. non-disclosers) can thus misestimate the true "treatment effect" of decarbonization policies unless the selection process is modeled.

Selection bias in carbon disclosure is amplified by the information constraints faced by data vendors. Vendors often lack granular data on smaller firms (e.g., limited public sustainability reporting and opaque supply chains), which makes imputations both necessary and potentially fragile. At the same time, vendors' methodologies are typically proprietary and not fully disclosed, making it difficult for researchers to assess which covariates drive peer matching and how missingness is handled. This motivates an explicit econometric treatment of selection and a transparent, data-driven approach to imputation. In this sense, selection bias in emissions is not merely an econometric nuisance but an economically meaningful distortion in a core input used for climate policy analysis (Hansen and Scheinkman (2025)).

A further difficulty is that the natural "sector" concept for emissions is activity-based and does not map cleanly into standard economic sector classifications. While the IPCC Guidelines set a clear method for differentiating between "sectors of economy" (Eggleston et al. (2006)), these sectors are groupings of activities, whereas in economics a sector is a grouping of similar economic actors. The energy sector, for example, includes most combustion of energy whether undertaken by energy producers or not; household gasoline combustion is counted in energy under IPCC, whereas under economic accounts it is included in the household sector. To align with the IPCC notion of sectors as activity-based groupings, we propose portfolio sorts as a method for grouping firms from the data: firms are sorted into L portfolios based on selected characteristics whose values reflect the activities undertaken by different groups of firms, thereby forming activity-based sectors.

These conceptual issues are mirrored in practice by substantial dispersion across vendor estimates. Carbon estimates provided by data vendors - including CDP, *Trucost*, MSCI, Sustainalytics, Thomson Reuters, Bloomberg, and ISS - often diverge due to differences in how "nearest peer groups" are defined. For instance, Thomson Reuters employs a carbon-to-size ratio to match potential peers, expanding from industry to broader classifications until sufficient observations exist, whereas MSCI uses total revenue rather than size as the primary matching criterion. Beyond size and revenue, many firm characteristics can proxy for operational scale and predict carbon output (e.g., assets, sales, employment, net PPE). These differences raise a fundamental question: which characteristics are most informative for identifying peer groups and inferring undisclosed carbon output, especially when disclosure is endogenous? Can we simply rely on naïve imputations based on size or revenue? The question is first-order in practice: research publications using vendor estimates by MSCI, Refinitiv, Sustainalytics, and Trucost have grown from a dozen per year in 2008 to several thousands in recent years, according to the Dimensions research database. In addition, adverse selection can be a policy instrument. Cicala et al. (2022) advocate a certification algorithm that policymakers can exploit to encourage certification by relatively clean firms based on voluntary disclosures when only the mean of the emissions distribution is known. Our selection-corrected imputation framework can be used to estimate that mean when it is unknown.

Our methodological contribution is to provide a transparent, high-dimensional correction for selection bias that is compatible with modern covariate-rich corporate datasets. We propose kernel group lasso (KG-lasso) to identify carbon neighbours across characteristics-sorted portfolios that mimic activity-based sectors. In brief, KG-lasso selects groups of portfolios sorted by key characteristics that are informative for carbon estimation, so that imputations are anchored in *carbon next door* rather than in ad hoc peer definitions. A kernel function produces weights over the L portfolios, measuring similarity between a non-disclosing firm and the portfolios in terms of selected characteristics. By leveraging a feature map, the kernel measures similarity in feature space rather than the original characteristics space, capturing nonlinearities and high-order interactions beyond simple linear correlations. For identification in the sample selection problem, we propose an adaptive KG-lasso that differentiates contributors to variable selection from contributors to sample selection: estimates from the selection equation provide adaptive weights that differentiate regularization at the group level, so that the active sets in the variable-selection and sample-selection equations have a bounded intersection. Ultimately, these adaptive weights support exclusion restrictions. Put differently, identification relies on high-dimensional exclusion through differential regularization rather than on classical low-dimensional instruments.

On the econometric side, we extend the two-step procedure of Heckman (1979) to a three-step pro-

cedure suitable for high-dimensional nuisance parameters.[4] The first step estimates high-dimensional nuisance components; the second step uses DML and Neyman orthogonality to consistently estimate the coefficient on the selection correction term and to construct a doubly robust score test for sample selection (Chernozhukov et al. (2018)); and the third step performs post variable selection in the outcome equation to obtain consistent estimation and selection. This focus yields two econometric outcomes: a doubly robust score test for detecting sample selection bias, and consistent variable selection in the presence of such bias. Our approach à la Heckman exploits a plug-in bias correction term and therefore differs from the control function approach of Newey et al. (1999) and Imbens and Newey (2009). Recent work by Bia et al. (2024), aligned with the control-function perspective, extends DML methods to sample selection models. In contrast, our emphasis is on (a) consistency of the bias-correction coefficient in a high-dimensional setting and (b) asymptotic normality for the post-selection estimators in the emissions equation.

We evaluate the quantitative importance of selection bias using annual carbon data from *Trucost* covering 3,444 unique U.S. firms and 22,043 firm-year observations over January 2010 to December 2023. A substantial share of these observations is vendor-estimated rather than firm-disclosed. We leverage 173 firm characteristics for both sample and variable selection. Our primary empirical result is a strong rejection of the null hypothesis of no sample-selection bias: the coefficient on the selection term is consistently negative and highly significant, indicating a negative correlation between unobserved determinants of disclosure and unobserved drivers of emissions. Ignoring selection bias leads to substantial underestimation of Scope 1, 2, and 3 emissions. Encouragingly, we document a steady decline in this bias over time, particularly for Scope 1 and 2; this trend appears related to expanded data coverage following acquisition of *Trucost* by S&P in 2016.

A second finding concerns the economics of selection and the informational content of covariates. Selection is primarily driven by indicators of firm quality, with size, age, and trading volume emerging as dominant predictors. In contrast, covariates selected in the emissions equation relate more to growth opportunities (R&D intensity, profitability, investment activity) and capital structure; notably, firm size plays no significant role in the variable-selection stage. These results underscore the empirical relevance of a high-dimensional approach in both sample and variable selection. They also clarify why naïve imputations can perform poorly: when imputations rely only on firm size or revenue, the magnitude of implied underestimation becomes several orders of magnitude larger than when using the full set of characteristics and correcting for selection. This is a key concern for empirical work studying carbon premia or carbon burdens that relies on a small number of characteristics (see, e.g., Aswani et al. (2024), Bolton and Kacperczyk (2021, 2023), Pastor et al. (2025), Zhang

[4]Sample selection issues can also be addressed via other methods targeting misspecification of conditional distributions (Chen et al. (2024)).

(2025)).

A third finding concerns the fiscal and welfare implications of underestimation. Using our selection-corrected high-dimensional approach and accounting for underestimation across all scopes, we estimate a conservative carbon tax revenue loss of $2.65 billion. This estimate is likely understated because the methodology used by data vendors to impute emissions is typically undisclosed. Comparing vendor-imputed emissions to our selection-corrected predictions, we observe substantial underestimation in vendor data as well. We do not claim that vendors neglect selection bias; rather, we show that underestimation can arise mechanically when disclosure is endogenous and imputations are constrained by incomplete information. This discrepancy translates into a potential tax revenue shortfall exceeding $9 billion. A forward-looking perspective can be gained by considering the social cost of carbon. The U.S. Environmental Protection Agency (EPA) periodically publishes estimates of this cost (see EPA (2023)), and Folini et al. (2025) show its sensitivity to model specification and calibration. Using our corrected emissions estimates, we find that the implied economic cost of underreported corporate emissions could be as high as $525 billion, far exceeding the estimated tax shortfall. We also consider a flexible parametric extension beyond normality based on the two-step series estimator of Newey (2009); results are qualitatively similar and underestimation magnitudes are close. More broadly, these magnitudes underscore that accurate emissions measurement is a key input into optimal climate policy and valuation (Cai and Lontzek (2019), Acemoglu et al. (2016)).

Toward financial markets, our results have implications for how climate risk is priced when the underlying emissions inputs are biased. A related literature emphasizes that climate valuation depends on risk, ambiguity, and model misspecification (Barnett et al. (2020)). Survey evidence also suggests that market participants perceive asset prices as underestimating climate risks (Stroebel and Wurgler (2021)). Our analysis provides an upstream mechanism consistent with this perception: when emissions of non-disclosers are understated due to naïve imputations and uncorrected selection, firms are misclassified between "low" and "high" emitter portfolios. Even if imperfect, there is solid evidence that vendor rankings and labels are taken seriously by investors (Hartzmark and Sussman (2019), Ceccarelli et al. (2024)). We show that selection-corrected emissions substantially alter rankings relative to vendor-, size-, and revenue-scaled imputations and imply a markedly larger carbon premium (in absolute value) than estimates obtained without controlling for selection.

Our findings also speak to the broader disclosure and regulation literature. The sharp contrast between expanding disclosure coverage and uneven disclosure quality highlights an information environment problem: even when emissions data exist, they can be noisy and unevenly reliable, exposing firms to information risk when disclosure is partial or absent. By providing more accurate estimates of emissions, improved measurement can strengthen firms' incentives to disclose, particularly when firms

anticipate that their true emissions differ from third-party estimates. More broadly, ESG disclosure shapes asset prices and capital allocation (Ilhan et al. (2023)), and long-term shareholder activism can induce firms to disclose climate-risk exposures and generate valuation premiums (Flammer et al. (2021)). At the same time, without uniform mandates, disclosure remains uneven, reflecting managerial uncertainty and intertemporal trade-offs (Bond and Zeng (2022), Kremer et al. (2024)). Recent regulatory innovations indicate that mandatory climate reporting can improve disclosure quality and redirect capital toward lower-carbon investments (Gibbons (2024), Gehricke et al. (2025)), while well-designed ESG mandates enhance stock liquidity, particularly for firms with weaker information environments (Krueger et al. (2024)). These findings suggest that better measurement and better disclosure are complementary.

We do not claim that improved measurement alone is a panacea for climate mispricing or policy design. Even when emissions are correctly measured and priced, external constraints and structural frictions can limit the effectiveness of market-based climate policies (Fowlie et al. (2016), Nath (2025), Lyubich (2025)). Our contribution is complementary: we document a foundational channel through which endogenous disclosure and sample selection distort emissions data upstream of pricing, policy design, and adaptation frictions.

The paper is organized as follows. In Section 2, we outline our model based on a high-dimensional regression with reproducing kernels, comprising a variable selection equation and a sample selection equation, and we discuss identification. We also build a doubly robust score test for our sample selection model, deploying the DML approach of Chernozhukov et al. (2018) to correct for regularisation bias. In Section 3, we show how kernel group lasso delivers consistency of variable selection under sample selection bias. In Section 4, we describe the data and empirical results. In the Appendix, we provide proofs of our theorems. In an Online Appendix, we gather additional Tables and Figures and provide an overview of reproducing kernel methods.

2 A high-dimensional model with variable selection and sample selection

2.1 High-dimensional regression with reproducing kernels

Our primary goal is to estimate carbon emissions using its carbon neighbours identified by kernel group lasso. Let us denote the disclosed carbon of firm i by Y_i, while $X_i = (X_{i1}, \cdots, X_{iJ})$ is a J

dimensional characteristics of i, and J is high-dimensional. We suppose that we are given n firms' information on carbon emissions. A prediction of Y_i given the carbon outputs of its neighbouring firms is $Y_i = \mathsf{E}\big[Y_k | k \in \mathcal{N}_i\big]$, so that we take the average of Y_k for those k considered as the neighbour of i denoted by $\mathcal{N}_i$.

How can we find neighbours that offer sufficient carbon insights? We can identify the potential neighbours using some firm characteristics to define neighbourhood. We can lay out the unknown carbon function conditional on firm characteristic information. To avoid the curse of dimensionality, like many nonparametric approaches, we impose an additive model to approximate the unknown carbon function. The additive property is the by-product of reproducing properties in the Reproducing Kernel Hilbert Space (RKHS) such that a linear combination of kernels is a kernel function per se.[5] Unlike the conventional nonparametric approach such as kernel smoothing that exploits a cross-section of $n-1$ to identify/weight potential neighbors, we identify a group of neighbors, that is, the portfolios sorted by one particular characteristics for which i may belong. A sorted-portfolio is a group of firms with similarity in characteristic j as sorting criteria. There are two main reasons for this. First, in the finance literature, portfolio-sorting is popular in return prediction as it flexibly takes advantage of cross-sectional information (Freyberger et al.; 2020). The second reason is that portfolio sorts allow us to form activity-based groupings. Firms are sorted into L portfolios based on the values of selected characteristics. These characteristics and their values are chosen to reflect the type of activities undertaken by different groups of firms, thereby forming activity-based sectors. An immediate benefit is that the computing load is reduced from $n(n-1)$ to $n \times L$, provided $L \ll n$.

In our empirics, Y_i is the firm-level carbon emissions in tons of carbon dioxide in logarithm. We can characterize the conditional mean equation in the RKHS as a linear span of reproducing kernels. As such, we can have a simple estimate if we can allocate i to portfolio ℓ sorted by the value of characteristic j, denoted as $P_{j,\ell}$ for $\ell = 1, \cdots, L$, $j = 1, \cdots, J$, namely

$$Y_i = \sum_{j=1}^{J} \sum_{\ell=1}^{L} b_{j,\ell} k(X_{ij}, P_{j,\ell}) + \epsilon_i, \tag{1}$$

where $P_{j,\ell}$ denotes the representative point of the ℓ-th portfolio (e.g., the midpoint of the characteristic range), serving as the second input argument for the kernel function. Here, $k(.,.)$ is a known reproducing kernel function for which we can choose from a family of kernel functions such as polynomial kernels or Gaussian kernels.[6] Gaussian kernel acts as a "catch-all" method as it never performs poorly than others (Exterkate (2013)). For this reason, we use Gaussian kernels in our

[5] See Berlinet and Thomas-Agnan (2011) for more details on RKHS and the review in section OA.2.
[6] See Appendix OA.2 for more details.

empirical exercise. The corresponding coefficient $b_{j,\ell}$ weights the kernel function. In the general additive specification of (1), $b_{j,\ell}$ is the "partial-regression coefficient" (or weight) associated with the feature $k\left(X_{ij}, P_{j,\ell}\right)$, representing the marginal contribution of characteristic j's ℓ-th portfolio to the total predicted emission Y_i, after accounting for the influence of all other $JL-1$ features. The simpler interpretation that $b_{j,\ell}$ is the average carbon of the ℓ-th portfolio sorted by j holds true only in the special case where a single characteristic j is used with non-overlapping kernel features (e.g., indicator functions $\mathbf{1}\left(X_{ij} \in I_{j,\ell}\right)$) where $I_{j,\ell}$ indicates an interval of the distribution for a given firm characteristic j. We can think of $k\left(X_{ij}, P_{j,\ell}\right)$ as a smooth extension of the indicator function $\mathbf{1}\left(X_{ij} \in I_{j,\ell}\right)$ to belong to $I_{j,\ell}$. Here, we would like to incorporate higher-order information from a set of variables including $X_{i,j}$, $P_{j,1}, ..., P_{j,L}$, and their nonlinear interaction. We discuss useful functional properties of kernels in the following subsection.

Let

$$\mathbf{k}(X_i) = \begin{pmatrix} k(X_{i1}, P_{1,1}), \cdots, k(X_{i1}, P_{1,L}) \\ \vdots \\ k(X_{iJ}, P_{J,1}), \cdots, k(X_{iJ}, P_{J,L}) \end{pmatrix}_{J \times L}. \tag{2}$$

Applying the vectorization operator $\mathsf{vec}(\cdot)$ that stacks the rows of $\mathbf{k}(X_i)$ on top of one another to yield $\mathbf{k}_i = \mathsf{vec}(\mathbf{k}(X_i)^\top)^\top$ and $\mathbf{k}_i \in \mathbb{R}^{1 \times JL}$. $\mathbf{k}_i$ represents the kernel evaluated at X_i. We obtain a compact representation of (1), namely

$$Y_i = \mathbf{k}_i b + \epsilon_i, \tag{3}$$

where $b \in \mathbb{R}^{JL \times 1}$ is the vector of parameters to be estimated. Hence, Equation (3) takes the form of a high-dimensional parametric linear regression based on reproducing kernels with many regressors $\mathbf{k}_i$ and many parameters in b (and as such, it should not be seen as a nonparametric estimation device). Indeed, in our empirics, we have $J = 173$ characteristics and $L = 10$ portfolios. In the spirit of Fama and French (1993) for factor construction based on deciles (see also Freyberger et al. (2020)), we set $L = 10$.[7] We may extend this ad-hoc choice to an adaptive or data-driven one to decide on L. Empirically, we find that the choice of L does not change much of our main findings.

[7] To construct their empirical factors, Fama and French (1993) sort stocks according to deciles of the firm characteristic. For example, for firm size, big stocks are those in the top 90% of June market cap, and small stocks are those in the bottom 10%.

2.2 Model setup

Estimating the conditional mean function and variable selection are unlikely to be consistent in the presence of sample selection bias if $\mathbf{k}_i b$ in (3) is estimated using non-randomly selected subsamples. It is the main challenge in a growing body of literature for carbon estimation. Carbon information Y_i is observed if firm i reports its emission estimate, indicated by $D_i = 1$, otherwise, Y_i is unknown due to $D_i = 0$. Let N denote the entire sample size, and use n to denote the subsample for which $D_i = 1$. The Variable Selection and Sample Selection (`VS-SS`) high-dimensional model that we study comprises two equations, one for variable selection in (4) that aims to select carbon-relevant characteristics from a full set of X and estimate the unknown carbon regression function using the n disclosed sample. The other one is for sample selection in (5) that models disclosure decision using full sample information N. The sample selection equation is the propensity score of disclosure conditional on regressors Z_i:

$$\texttt{variable selection:} \qquad Y_i = \mathbf{k}_i b + \epsilon_i, \tag{4}$$

$$\texttt{sample selection:} \qquad D_i = \mathbf{1}\big(Z_i\beta + \mathsf{v}_i > 0\big), \tag{5}$$

where $\mathbf{1}(\cdot)$ is an indicator function, which takes the value one if its argument is true and zero otherwise, $b \in \mathbb{R}^{JL\times 1}$ and $\beta \in \mathbb{R}^{p\times 1}$. The selection outcome D_i is endogenous, raising a selection bias due to

$$\mathsf{E}\big[\epsilon_i\big|\mathsf{v}_i, D_i = 1\big] \neq 0. \tag{6}$$

In the presence of sample selection bias, b in (4) cannot be consistently estimated using the observed sample $D_i = 1$, because $\mathsf{E}\big[Y_i\big|X_i, Z_i, D_i = 1\big] = \mathbf{k}_i b + \mathsf{E}\big[\epsilon_i\big|X_i, Z_i, D_i = 1\big] \neq \mathbf{k}_i b$ from (6). Given that the hidden selection takes a linear form $Z_i\beta$, and if we assume ϵ_i and v_i are bivariate normal random variables, Heckman (1979) shows that

$$\mathsf{E}\big[\epsilon_i\big|X_i, Z_i, D_i = 1\big] = \theta h_i, \tag{7}$$

where $h_i := h(-Z_i\beta)$, $h(z) := \phi(z)/\Phi(-z)$ is known as the inverse Mills ratio (IMR), and θ is proportional to the covariance between ϵ and v, denoted by $\sigma_{\epsilon,\mathsf{v}}$. For consistency in carbon estimation, it is stringent to incorporate (7) into (4) for bias correction. We rewrite the primary equation for the disclosed samples as

$$Y_i = \mathbf{k}_i b + \theta h_i + \varepsilon_i. \tag{8}$$

The conventional Heckman approaches, either one-step or two-step procedures, appear to collapse in the presence of high-dimensional variables in both sample selection and variable selection equations.

It is stringent to consider advanced modern approaches to circumvent the curse of dimensionality in the `VS-SS` framework. For this, we consider DML proposed by Chernozhukov et al. (2018), which leverages the Neyman orthogonality to make θ insensitive to inconsistency in the high-dimensional nuisance estimates b and β. The inconsistency arises from the regularisation bias in the nuisance estimates. As long as θ, the coefficient of selection bias, can be consistently estimated, one can do "post" variable selection in the variable selection equation to attain consistent estimators and consistent variable selection. For consistency of variable selection, we extend a two-step procedure of Heckman (1979) to a three-step procedure. The first step is to estimate nuisance parameters b and β, and plug in these nuisances into (8) to consistently estimate θ in the second step using DML approach. In the last step for post variable selection, we consistently estimate b and derive consistent variable selection for an active subset of b as desired. This three-step procedure generalises Heckman (1979) to a high-dimensional framework. In addition to estimation strategies, we establish asymptotic analysis in the `VS-SS` framework. The asymptotic analysis in the proposed framework decouples from Heckman (1979) because jointly asymptotically analysing θ, β and b is not possible in the presence of the curse of dimensionality. The estimators of β and b suffer potentially from bias caused by regularisation. We need to first asymptotically analyse $\hat{\theta}$, $\hat{\beta}$ separately, then given the studied asymptotic properties of $\hat{\theta}$ and $\hat{\beta}$ we finally derive an asymptotic analysis on $\hat{b}$ for estimation consistency and variable selection consistency.

We begin with estimating nuisance parameters as the first step. The selection equation in (5) under the normally distributed error terms can be parametrised by lasso probit that allows us to handle high-dimensional Z_i and undertake variable selection in the parametrised propensity score $Z_i\beta$. $\beta \in \mathbb{R}^p$, a p-dimension vector, can be estimated by lasso probit to penalise small values in the parameter vector. The linear form with a sparsity constraint facilitates understanding of important covariates that determine the propensity of disclosure:

$$\hat{\beta} = \arg\min_{\beta} \mathsf{E}_N\left[\Lambda_i(\beta)\right] + \lambda_1 \left\|\beta\right\|_1, \tag{9}$$

where E_N denotes the sample mean of N observations, $\Lambda_i(.)$ is the negative log-likelihood for the probit model evaluated at i, and $\|\cdot\|_1$ is L^1-norm driven by the penalisation parameter $\lambda_1 > 0$.

Now, we turn to the estimation of b, the coefficient used to weight kernel functions in the kernel group lasso model. This estimate is informative as it sheds light on the identification of carbon neighbours across characteristic-sorted portfolios that mimic activity-based sectors. If we think that characteristic j is the most informative to identify the carbon neighbours, then the reproducing kernel k_j spans the RKHS. However, we may desire sparsity in the sense that only a small subset of characteristics spans the RKHS. Therefore, we impose a regularisation for complexity as follows with

penalisation parameter $\lambda > 0$:

$$\min_{b} \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - \mathbf{k}_i b \right)^2 + \lambda \sum_{j=1}^{J} \|b_j\|_{\mathsf{K}_j}, \tag{10}$$

where $b \in \mathbb{R}^{JL \times 1}$ stacks all column vector b_j for $j = 1, ...J$, and $b_j \in \mathbb{R}^{L \times 1}$ a L-vector of coefficients. In (10), the regularisation factor corresponding to j is $\|b_j\|_{\mathsf{K}_j} = \left(b_j^\top \mathsf{K}_j b_j\right)^{1/2}$, where K_j is a symmetric $L \times L$ positive definite kernel matrix, with entries $\left[\mathsf{K}_j\right]_{\ell,\ell'} = \frac{1}{L} k(P_{j,\ell}, P_{j,\ell'})$, for $\ell, \ell' = 1, \cdots, L$. We choose Gaussian basis kernels in the empirical study. The good properties of Gaussian kernels have been discussed in the online appendix. In the penalty term, the coefficient vector b_j is weighted by the kernel matrix K_j. The norm $\|b_j\|_{\mathsf{K}_j}$ in the RKHS space has salient insights in terms of cross-sectional information across L portfolios sorted by a given characteristic. The entry $\left[\mathsf{K}_j\right]_{\ell,\ell'} = k(P_{j,\ell}, P_{j,\ell'})$ for all $\ell, \ell' \in L$ measures the similarity between portfolios ℓ and ℓ', sorted by j. If all entries in K_j are large or close to one, it implies a general large similarity among the L portfolios sorted by j. In other words, characteristic j, as sorting criteria, is not discriminant enough along the cross-section of firms grouped into the L portfolios. It explains why we want to penalize the coefficients associated to such an uninformative characteristic through a large weight in $\|b_j\|_{\mathsf{K}_j}$ in order to discard j in building activity-based sectors. The penalty function in a group lasso is intermediate between the L^1-penalty that is used in the lasso and the L^2-penalty that is used in ridge regression (see Yuan and Lin (2006) for graphical illustrations of the different penalties). The group lasso encourages sparsity at the group level, and not within individual groups.

The solution for $b \in \mathbb{R}^{JL \times 1}$ should be sparse. We can re-express (10) as follows:

$$\min_{b} \frac{1}{2} \|Y - \mathbf{k}b\|^2 + \lambda \sum_{j=1}^{J} \|b_j\|_{\mathsf{K}_j}, \tag{11}$$

where $\mathbf{k} \in \mathbb{R}^{n \times JL}$ stacks n kernel vectors $\mathbf{k}_i$, $i = 1, \cdots, n$. Denote $\boldsymbol{z}_j = \mathbf{k}(j)^\top \left(Y - \mathbf{k} b_{-j}\right)$, $b_{-j} = \left(b_1, \cdots, b_{j-1}, \mathbf{0}, b_{j+1}, \cdots, b_J\right)^\top$ and

$$\mathbf{k}(j) = \begin{pmatrix} k(X_{1j}, P_{j,1}), \cdots, k(X_{1j}, P_{j,L}) \\ \vdots \\ k(X_{nj}, P_{j,1}), \cdots, k(X_{nj}, P_{j,L}) \end{pmatrix}_{n \times L}. \tag{12}$$

Following Yuan and Lin (2006), we assume that each $\mathbf{k}(j)$ is orthonormalised on the blocks of the design matrix $\mathbf{k}$. This assumption is standard practice in many large-scale Group Lasso implementations, as it ensures the Lipschitz constant of the quadratic loss function with respect to b_j is

approximately unity. Let $\mathbf{I}_L$ be L-dimensional matrix of ones. A closed-form solution for (11) is given by the proximal operator of the group Lasso penalty

$$\hat{b}_j = \left(\mathbf{I}_L - \frac{\lambda \mathsf{K}_j}{\|\boldsymbol{z}_j\|_{\mathsf{K}_j}} \right)_+ \boldsymbol{z}_j. \tag{13}$$

The positive-part operator $(\cdot)_+$ denotes the matrix obtained by spectral truncation, where all negative eigenvalues of the matrix argument are replaced with zero.

2.3 Identification issues in a high-dimensional system

It is crucial to consider identification issues in (8). In most cases, Z_i and X_i will have many variables in common. A strong form of exclusion restrictions is not possible in the high-dimensional case. In the case of the sample selection model, in order to separately identify the decision regarding disclosure (to report or not to report) from the carbon determinants (how much to emit carbon), we must have variables that enter Z_i but do not enter X_i. If such variables (known as exclusion restrictions) cannot be found then separate identification depends upon the nonlinearity of the extra term (known as the IMR) which appears in the variable selection equation. However, as addressed by Vella (1998), the IMRs are likely to be linear over a wide range of their arguments.

To ensure identification and facilitate estimation and variable selection in a high-dimensional setting, we rely on the following assumptions.

Assumption 1. *ϵ_i and v_i are i.i.d. jointly normally distributed with zero means and finite second moment σ^2 and covariance $\sigma_{\epsilon,\mathsf{v}} \neq 0$, and $(\epsilon_i, \mathsf{v}_i)$ are independent of Z_i.*

Assumption 2. *The exclusion restrictions require $supp(X_i) \subset supp(Z_i)$ and X_i is contained in Z_i.*

Assumption 3. *Let $\mathsf{A}_\beta = \{j; \beta_j \neq 0\}$ be the active set of selection parameters, and let $\mathsf{A}_b = \{j; b_j \neq 0\}$ be the active set of b determining carbon emissions. Let $|\mathsf{A}_b|$ and $|\mathsf{A}_\beta|$ be cardinality of A_b and A_β. As sparsity condition, we assume $|\mathsf{A}_b| < J$ and $|\mathsf{A}_\beta| < q$ and $q < p$.*

Assumption 4. *Denote P^d the power set of $\{1, \cdots, d\}$ and $d = \min(|\mathsf{A}_b|, |\mathsf{A}_\beta|)$. For an intersecting family $\mathcal{A} \subseteq \mathsf{P}^d$, and $1 \leq s \leq n$ define an intersection structure of $\mathcal{A}$ by $\mathbf{I}(\mathcal{A}) = \{\mathsf{A}_\beta \cap \mathsf{A}_b : \mathsf{A}_\beta, \mathsf{A}_b \in \mathcal{A}\}$ and the collection of s-intersections of $\mathcal{A}$ by $\mathcal{A}(s) = \{\mathsf{A} \in \mathbf{I}(\mathcal{A}) : |\mathsf{A}| = s\}$.*

Assumption 1 and 2 are primitive. Assumption 1 enables the adoption of the IMR to achieve bias correction by plug-in. Assumption 2 is a mild identification condition. The exclusion restrictions

are stated in terms of X_i and not their transformations $\mathbf{k}_i$, since it is the behavior of the former that matters in terms of identification. Assumption 3 allows Z_i and X_i to share many variables, and their active sets are small relative to their corresponding size of the full set. In Assumption 4, we impose that two active sets are nearly disjoint. For this, we require that the two active sets have a bounded intersection.[8] Assumption 4 suffices to ensure nonlinearity of IMR, at least over part of the range of its arguments, even if the two design matrices have considerably overlapping columns. To meet these assumptions, we introduce an adaptive version of kernel group lasso. The estimates from the selection equation underlie adaptive weights used to differentiate regularisations at the group level. Intuitively, if $X_j = Z_{j\in \mathsf{A}_\beta}$, we want to penalize its associated coefficient b_j more heavily with a regularisation directly weighted by $(1+|\hat{\beta}_j|)^\gamma$, for $\gamma \geq 1$. To get that X_i is contained in Z_i in Assumption 2, we let $Z = (X, U)$ and $Z_j = X_j$ for $j = 1, \cdots, J$ and $Z_j = U_{j'}$ for $j = J+1, \cdots, p$ and $j' = 1, \cdots, (p-J)$. Then, we build an adaptive Kernel group lasso:

$$\min_b \frac{1}{2} \|Y - \mathbf{k}b\|^2 + \lambda \sum_{j=1}^{J} w_j \|b_j\|_{\mathsf{K}_j}, \tag{14}$$

where $w_j = \big(1+|\hat{\beta}_j|\big)^\gamma$, $\gamma \geq 1$ for $j = 1, \cdots, J$ and $\hat{\beta}_j$ is the estimate from (9). The magnitude of $|\hat{\beta}_j|$ determines an additional weight in excess of one. Clearly, the adaptive-weight regularisation in (14) boils down to the plain regularisation with $w_j = 1$ if $\hat{\beta}_j = 0$. Under such an adaptive weighting scheme, Assumptions 3 and 4 are satisfied.

To rigorously pin down the asymptotic theorem for θ and to test sample selection bias, we incorporate Neyman orthogonality conditions and derive the Neyman orthogonal score to make θ insensitive to inconsistency in the plug-in estimates. The resulting estimator is M-estimator. The idea of adopting Neyman orthogonal score estimation can be dated back to Newey (1994). Newey (1994) gives conditions on estimating equations and nuisance function estimators so that nuisance function estimators do not affect the limiting distribution of parameters of interest. Chernozhukov et al. (2018) establishes the equivalence between Neyman orthogonal score and the partialling-out approach of Robinson (1988).

Let $\mathcal{W} = \{W_i\}_{i=1,\cdots,N}$ be a collection of $W_i = \big(Y_i, D_i, Z_i, X_i\big)$. Suppose $\ell(W;\theta,\eta)$ a known criterion function which is continuously differentiable almost surely, $\ell(\cdot)$ can be a log-likelihood or quasi-log-likelihood function. In the case of a linear model, we have $\ell\big(W_i;\theta,\eta\big) = -\frac{1}{2}\big(Y - \mathbf{k}_i b - \theta h_i\big)^2$. We present the Neyman orthogonal score $\psi(W;\theta,\eta) = \partial_\theta \ell(W;\theta,b) - \mu \partial_b \ell(W;\theta,b)\big)$, where $\eta = \big(b,h\big)$ and $h := h(Z\beta)$. μ solves the equation $\mathcal{J}_{\theta\eta} - \mu \mathcal{J}_{\eta\eta} = 0$, where $\mathcal{J}$ stands for Jacobian

[8] Assumption 4 is compatible with the Erdös-Ko-Rado Theorem, which limits the number of sets in a family of sets for which every two sets have at least one element in common.

matrix. $\mu_0 = \mathcal{J}_{\theta\eta}(\mathcal{J}_{\eta\eta})^{-1}$ is a unique solution if $\mathcal{J}_{\eta\eta}$ is invertible. The orthogonality conditions imply $\partial_\eta \mathsf{E}\big[\psi(W;\theta_0,\eta_0)\big][\eta-\eta_0]=0$. With the sample selection function $h(Z\beta)$ and the variable selection function $\mathbf{k}b$ being parametrised in a high-dimensional setup, the Neyman orthogonal score for the `VS-SS` model is given by

$$\psi(W_i;\theta,\eta) = \big(Y_i - \mathbf{k}_i b - \theta h_i\big)\big(h_i - \mathsf{E}[h|X_i]\big). \tag{15}$$

Equation (15) forms a linear score (linear in θ) to benefit from computational advantages to side step Jacobin matrix computation. We can decompose the entire score into two parts, $\psi(W_i;\theta,\eta) = \psi_a(W_i;\eta)\theta + \psi_b(W_i;\eta)$,

where $\psi_a(W_i;\eta) = -h_i(h_i - \mathsf{E}[h|X_i])$ and $\psi_b(W_i;\eta) = (Y_i - \mathbf{k}_i b)\big(h_i - \mathsf{E}[h|X_i]\big)$. More explicitly, the linear score implies $\mathbb{E}[\mathbf{k}_i h_i] = 0$ after centering. For $\theta = \theta_0$, we have a moment condition satisfying

$$\mathsf{E}\big[\psi(W_i;\theta_0,\eta_0)\big] = 0. \tag{16}$$

Assumption 5 (Linear score with approximate Neyman orthogonality). *There exists a twice continuously Gateaux-differentiable map $\eta \to \mathsf{E}\big[\psi(W_i;\theta_0,\eta_0)\big]$, and the score ψ satisfies the Neyman orthogonality: $\partial_\eta\psi(\theta_0,\eta_0)[\eta-\eta_0]=0$.*

Algorithm 1: Estimating selection bias via K-fold cross-fitting

Input: $\mathcal{W} = \{W_i\}_{i=1,\cdots,N}$ be a collection of $W_i = \big(Y_i, D_i, Z_i, X_i\big)$

1 Split $\mathcal{W}$ in K subsamples. For each subsample k, let n_k be its size, $\mathcal{W}_k$ be k-th fold subsample, and $\mathcal{W}_k^c$ be its complement set.

2 Split $\mathcal{W}_k^c$ into 2 nonoverlapping subsamples and estimate the nuisance parameter β in one subsample, and b in the other subsample to predict β_k and b_k in $\mathcal{W}_k$.

3 Estimate θ_k using the plug-in $\eta_k = \{\beta_k, b_k\}$ under the moment condition in (16).

4 Average θ_k across all K subsamples to obtain $\hat{\theta} = \frac{1}{K}\sum_{k=1}^{K}\theta_k$.

Output: $\hat{\theta}$

To satisfy the orthogonality conditions, following Chernozhukov et al. (2018) we deploy the K-fold cross-fitting Algorithm 1 to estimate $\eta = (\beta, b)$. Such random sample splitting avoids overfitting issues in nuisances and mitigates high variance caused by the use of subsamples. The estimate of η from auxiliary samples will be plugged into the score function to solve the moment condition for θ estimate. Given the efficient orthogonal score, $\hat{\theta}$, the average of θ estimates from K folds, is the consistent estimator of θ_0, as stated in the next theorem following directly from Chernozhukov et al. (2018) and based on the implementation algorithm 1. Let us split $\mathcal{W}_k^c$ into 2 non-overlapping

subsamples and estimate the nuisance parameter β in one subsample, and b in the other subsample to predict β_k and b_k in $\mathcal{W}_k$. This second split provides an additional safeguard against potential dependence or correlation between the two high-dimensional nuisance estimators ($\hat{\beta}$ and $\hat{b}$), and it is a robustness refinement that can be omitted in favor of using the full complement sample $\mathcal{W}_k^c$ for both nuisance fits.

Theorem 1. *If the score is the Neyman orthogonal efficient and satisfies Assumption 5, and under the condition that the variance of the score ψ is non-degenerate: all eigenvalues of matrix $\mathsf{E}\big[\psi\big(W;\theta_0,\eta_0\big)\psi(W;\theta_0,\eta_0\big)^\top\big]$ are bounded from below by $c>0$, then we have*

$$\sqrt{N}\big(\hat{\theta}-\theta_0\big)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\bar{\psi}(W_i)+O_p(\tau_N)\to\mathcal{N}(0,\sigma_\psi^2), \tag{17}$$

where the remainder term requires $\tau_N=o(1)$ with τ_N subsuming the estimation errors from all nuisance parameters in their second order terms (Chernozhukov et al.; 2018), and $\bar{\psi}(W_i):=-\mathcal{J}_0^{-1}\psi(W_i,\theta_0,\eta_0)$ is the influence function, and Jacobian matrix $\mathcal{J}_0:=\partial_\theta\mathsf{E}\big[\psi(W,\theta,\eta_0)\big]\big|_{\theta=\theta_0}$, while the asymptotic variance is $\sigma_\psi^2:=\mathcal{J}_0^{-1}\mathsf{E}\big[\psi(W,\theta_0,\eta_0)\psi(W,\theta_0,\eta_0)^\top\big](\mathcal{J}_0^{-1})^\top$.

We can replace σ_ψ^2 by a consistent estimator $\hat{\sigma}_\psi^2$, obtained via the K-fold cross-fitting using Algorithm 1 (see Theorem 3.2 of Chernozhukov et al. (2018)). The risk bound $O_p(\tau_N)$ in (17) embeds the risk arising from (i) the deviation from the orthogonal condition; (ii) the smoothness of the score function; (iii) the estimation risk of the Jacobian matrix; (iv) the risk associated with the estimated score. The estimation risk of nuisance functions including β, b is at the conventional nonparametric rate $o(N^{-1/4})$, as a result, their product rate is well-behaved at an approximate rate $O(N^{-1/2})$. Regarding the estimation risk of Kernel group lasso, Theorem 2 of Yang et al. (2016) formalizes, with a high probability, the estimation risk is $O(N^{-1/4})$. For further details on the regularity conditions and the proof of Theorem 1, we refer to Chernozhukov et al. (2018) and Newey (1994), and do not repeat them explicitly here.

We take a step further to conduct the score test based on the Neyman orthogonal score. The existing score test or LM test are limited in the presence of high-dimensional covariates in the variable selection equation. These existing tests fail to have unit power asymptotically against a wide range of regularisation bias in the variable selection equation. In addition, the existing tests may have low power in the presence of regularisation bias in the nuisance parameters. This low power is caused by the possibility that the bias term dominates the variance of the score. As a result, bias potentially governs the limiting distribution of test statistics, misleading the testing results (Hong and White; 1995; Wooldridge; 1992). To isolate the impact from bias, we undertake a doubly robust score test that replaces the efficient score vector with the Neyman orthogonal score. Under $H_0:\theta=0$, define

the sample mean score $\bar{\psi}_n(0,\hat{\eta}) = \frac{1}{n}\sum_{i=1}^n \psi_i(0,\hat{\eta})$, where $\hat{\eta}$ is obtained from the cross-fitted first-stage estimators described in Algorithm 1. Let $\hat{\sigma}_\psi^2 = \frac{1}{n}\sum_{i=1}^n \left(\psi_i(0,\hat{\eta}) - \bar{\psi}_n(0,\hat{\eta})\right)^2$ be the corresponding variance estimator (also cross-fitted). The doubly robust orthogonal-score test statistic is

$$s_n = \frac{n\,\bar{\psi}_n(0,\hat{\eta})^2}{\hat{\sigma}_\psi^2}. \tag{18}$$

Theorem 1 implies that LM-type statistics in (18) follows $s_n \xrightarrow{d} \chi_1^2$. In (18), we evaluate the score at the restricted model with $\theta = 0$ and $b = \hat{b}$. We do not reject the null hypothesis if s_n is sufficiently near zero. The test statistic s_n is a quadratic form of the score function, which is a linear function of θ. The proposed statistic is a standard LM (score) test constructed from a Neyman-orthogonal moment. Its asymptotic behavior and local power properties follow from classical score-test theory, (Newey; 1994). Our paper does not introduce a new testing principle but applies this well-established score-test framework in a high-dimensional sample-selection setting. Since the focus is on identification and bias correction rather than test comparison, we do not include a simulation study.

3 Variable selection under sample selection bias

3.1 Post variable selection consistency

In the presence of sample selection bias confirmed by the doubly robust score test, the estimate of b and variable selection in (4) are by no means consistent. Correcting such bias is essential for achieving consistency. It is clear that we are no longer relying on (13) for variable selection, and a necessary modification is introduced.

Let $\boldsymbol{z}_j(\theta)$ be function of θ for $j = 1, ..., J$, $\boldsymbol{z}_j(\theta) = \mathbf{k}(j)^\top\big(Y - \mathbf{k}b_{-j} - \theta h\big)$. The vector b_{-j} is defined as $(b_1, \ldots, b_{j-1}, \mathbf{0}, b_{j+1}, \ldots, b_J)^\top$, and $\mathbf{k}(j)$ is defined in (12). Having a consistent estimator of θ in Theorem 1, we obtain a post-selection estimator in the presence of sample selection bias

$$\hat{b}_j(\hat{\theta}) = \left(\mathbf{I}_L - \lambda w_j \frac{\mathsf{K}_j}{\sqrt{\boldsymbol{z}_j(\hat{\theta})^\top \mathsf{K}_j \boldsymbol{z}_j(\hat{\theta})}}\right)_+ \boldsymbol{z}_j(\hat{\theta}). \tag{19}$$

Equation (19) fundamentally deviates from (13) unless $\theta = 0$ under the null. Without a bias correction, the estimate of b may be overstated given the same value of λ. As discussed before, the adaptive weight w_j facilitates the implementation of identification. Invoking the Karush-Kuhn-Tucker conditions, we present the following proposition for post variable selection.

Denote $\|\mathbf{v}\|_{\mathbf{K}_j,*}$ the dual norm associated with the penalty norm $\|\cdot\|_{\mathsf{K}_j}$. Assuming K_j is invertible, the dual norm is defined as: $\|\mathbf{v}\|_{\mathsf{K}_j,*} = (\mathbf{v}^\top \mathsf{K}_j^{-1}\mathbf{v})^{1/2}$.

Proposition 1. *Let K_j with entries $\left[\mathsf{K}_j\right]_{\ell,\ell'} = \frac{1}{L}k(P_{j,\ell}, P_{j,\ell'})$ for $\ell, \ell' = 1, \cdots, L$ and for $j = 1, \cdots, J$. $\hat{\theta}$ is a consistent estimator of θ in Theorem 1. For identification purposes, the adaptive weight is given by $w_j = \left(1 + |\hat{\beta}_j|\right)^\gamma$, $\gamma \geq 1$ and $\hat{\beta}_j$ is the estimate from (9). Define $\|b_j\|_{\mathsf{K}_j} = \left(b_j^\top \mathsf{K}_j b_j\right)^{1/2}$, a regularisation factor corresponding to j. A necessary and sufficient condition for $\hat{b} \equiv \hat{b}(\hat{\theta}) = \left(\hat{b}_1(\hat{\theta}), \cdots, \hat{b}_J(\hat{\theta})\right)$ to be a solution is $-\mathbf{k}(j)^\top\left(Y - \mathbf{k}\hat{b}\right) + \lambda w_j \frac{\mathsf{K}_j \hat{b}_j}{\|\hat{b}_j\|_{\mathsf{K}_j}} = \mathbf{0}$, $\forall \hat{b}_j \neq \mathbf{0}$, and $\|\mathbf{k}(j)^\top(Y - \mathbf{k}\hat{\mathbf{b}})\|_{\mathsf{K}_j,*} \leq \lambda w_j$, $\forall \hat{b}_j = \mathbf{0}$.*

To ensure that Proposition 1 is plausible for variable selection, we establish the asymptotic properties of model selection to understand in which conditions it undertakes consistent model selection, when we are given a sufficient amount of sample size, namely $n \to \infty$. Let us establish the asymptotic properties for the selection consistency of the characteristics. The first step is to propose regularity conditions. Under these conditions, the estimates of the contributions of the characteristics to carbon estimation are non-zero for the characteristics that are truly relevant and shrink to zero for the irrelevant ones.

Regularity conditions We consider the case where $J > n$, meaning the dimensionality of firm characteristics (J) exceeds the cross-sectional sample size (n). This scenario is particularly likely during earlier periods, when the number of firm characteristics may surpass the number of firms with disclosed carbon information. We assume a diverging dimension in the sense that J_n can grow with n. Let true parameter be $b_0 = \left(b_{10}, \cdots, b_{J0}\right)$. Consistency in characteristics selection ensures a separation between the oracle active set $\mathsf{A}_b = \left\{j : b_{j0} \neq 0\right\}$ and the complement $\mathsf{A}_b^c = \left\{j : b_{j0} = 0\right\}$. We write $b_0 = \left(b_{10}', b_{20}'\right)'$, two subsets where b_{10} is the subset with the indices of elements in A_b and b_{20} is the subset with the indices of elements in A_b^c. The size of A_b is quantified by its cardinality $|\mathsf{A}_b| = J_1$ and by $|\mathsf{A}_b^c| = J_2$ for A_b^c, and $J_1 + J_2 = J_n$. For notational simplification, we suppress the subscript n in J_1 and J_2 for which both diverge on n. Define a submatrix of $\mathbf{k}$ to be $\mathbf{k}_1 \in \mathbb{R}^{n \times J_1 L}$ as we only consider the submatrix involving $j \in \mathsf{A}_b$. For each $j \in \mathsf{A}_b$ and the corresponding kernel matrix $\mathbf{k}(j)$, we denote a $L \times L$ covariance matrix $\Sigma_j = \frac{1}{n}\mathbf{k}(j)^\top \mathbf{k}(j)$. Likewise, we denote a $L \times L$ covariance matrix $\Sigma_{jh} = \frac{1}{n}\mathbf{k}(j)^\top \mathbf{k}(h)$ for $h \in \mathsf{A}_b$ and $j \notin \mathsf{A}_b$.

To present that the oracle properties of model selection consistency are attainable, we impose mild regularity conditions for the regularised model.

Condition 1 (Eigenvalue constraint). *Denote $\lambda_{min}(\mathcal{C})$ and $\lambda_{max}(\mathcal{C})$ as the minimum and maximum eigenvalue of a positive definite matrix $\mathcal{C}$. With $D_n > d_n \geq 0$ and $D_n < \infty$, it requires that, for*

each $j \in \mathsf{A}_b$, $d_n \leq \lambda_{min}(\Sigma_j) \leq \lambda_{max}(\Sigma_j) \leq D_n$.

Condition 2 (**Non-zero coefficients**). *The number of non-zero coefficients grows proportionally to cross-sectional sample size* n, *provided* $J_1 = \mathcal{O}(\log n)$

Condition 3 (**Partial orthogonality**). *Define the maximal carbon insight derived from any* $j \notin \mathsf{A}_b$, $c_n = \frac{\max_{j \notin \mathsf{A}_b} \left\| \mathbf{k}(j)^\top Y \right\|}{\left(\sum_{h \in \mathsf{A}_b} \left\| \mathbf{k}(h)^\top Y \right\| \right)/J_1} < 1$. *Partial orthogonality implies* $n^{-1}\lambda_{max}(\Sigma_{jh}) \leq \rho_n, \quad j \notin \mathsf{A}_b, h \in \mathsf{A}_b$. *For any* $\eta < 1$, *we assume* $\rho_n \leq \frac{d_n}{c_n\sqrt{J_1}}\eta$.

Condition 4 (**Irrepresentable condition**). *Denote a weighted sign vector* $\mathbf{s}_h = w_h \operatorname{sgn}(b_h)$ *for* $h \in \mathsf{A}_b$ *and an* L_2 *norm* $\|\mathbf{s}_h\|$. *For* $j \notin \mathsf{A}_b$ *and* $\mathbf{k}(h)^\top \mathbf{k}(h') = 0$ *for* $h \neq h'$, *the irrepresentable condition requires* $\frac{1}{n\sqrt{L}} \sum_{h \in \mathsf{A}_b} \left\| \mathbf{k}(j)^\top \mathbf{k}(h) \Sigma_h^{-1} \mathbf{s}_h \right\| \leq \frac{c_n \rho_n}{d_n} \leq \eta$. *The vector* $\operatorname{sgn}(b_h)$ *is defined as the scaled coefficient vector using the kernel metric:* $\operatorname{sgn}(b_h) = \frac{\mathsf{K}_h b_h}{\|b_h\|_{\mathsf{K}_h}}$, *where* $\|b_h\|_{\mathsf{K}_h} = \sqrt{b_h^\top \mathsf{K}_h b_h}$. *This vector* $\operatorname{sgn}(b_h)$ *is an element of the dual unit ball associated with the penalty and is the required subgradient component for active groups.*

We also rely on the fact that the off-diagonal cross-block terms in $\mathbf{\Sigma}_{\mathbf{k}_1}^{-1}$ are bounded and sufficiently small (i.e., $o(1)$) such that the expression is dominated by the diagonal blocks. Therefore, we use the inequality $\left\| n^{-1} \mathbf{k}(j)^\top \mathbf{k}_1 \Sigma_{\mathbf{k}_1}^{-1} \mathbf{s}_1 \right\| \leq \sum_{h \in \mathsf{A}_b} \left\| n^{-1} \mathbf{k}(j)^\top \mathbf{k}(h) \Sigma_h^{-1} \mathbf{s}_h \right\| + o(1)$ in our bounding argument. The condition on the maximum of this term ensures the irrepresentable condition holds for the inactive groups $\mathbf{b}_j = 0$.

Condition 5 (**Bounded eigenvalue of kernel matrix**). K_j *is a symmetric* $L \times L$ *positive definite kernel matrix, with entries* $\left[\mathsf{K}_j\right]_{\ell,\ell'} = \frac{1}{L} k(P_{j,\ell}, P_{j,\ell'})$ *for* $\ell, \ell' = 1, \cdots, L$, $j = 1, ..., J$. *With* $G_n > g_n \geq 0$ *and* $G_n < \infty$, *it requires for each* j, $g_n \leq \lambda_{min}(\mathsf{K}_j) \leq \lambda_{max}(\mathsf{K}_j) \leq G_n$.

Condition 1 requires the minimal eigenvalue for the covariance matrix of the kernel j, $\mathbf{k}(j) \in \mathbb{R}^{n \times L}$. It implies a covariance bounded away from zero such that the structure of similarity covariance of j-th feature among n firms is of low rank. Condition 2 controls the number of non-zero coefficients that grows proportionally to the cross-sectional sample size n. Condition 3 is a weak partial orthogonality assumption. It postulates a weaker correlation ρ_n between the reproducing kernel $\mathbf{k}_j$ from the active set and any one from its complement. It implies that the two subspaces, spanned by the reproducing kernels induced by the respective coordinates from the active and inactive sets, have a constrained intersection. Here, ρ_n is inversely bounded by c_n, the maximal carbon insight derived from any $j \notin \mathsf{A}_b$, to ensure that the correlation between Y and the irrelevant characteristics remains low relative to that with the relevant cluster. Condition 4 is the key to promoting model selection consistency. While much existing research (Chatterjee and Lahiri (2013), Huang, Horowitz and Ma (2008), Huang, Ma and Zhang (2008)) have relaxed a restricted assumption for complete orthogonality in the design

matrix, it remains unclear how to find an optimal bound for partial orthogonality. We establish the adaptive irrepresentable condition that incorporates maximal carbon relevance derived from the non-relevant characteristics for an optimal upper bound. It turns out that the structured upper bound is a necessary condition for selection consistency.

Selection consistency We opt for a strong form of selection consistency, namely sign consistency. It not only requires correctly distinguishing between a zero coefficient (inactive or irrelevant characteristics) and non-zero coefficients (active or relevant characteristics) but also ensuring that the signs of the estimated coefficients match those of the true coefficients. To formalize this, we introduce the sign equivalence operator $=_s$ to indicate $\hat{b} =_s b_0$ if $\text{sgn}(\hat{b}) = \text{sgn}(b_0)$.

Theorem 2 (**selection consistency of characteristics**). *If Conditions 1 - 5 hold, the adaptive KG-lasso selects the relevant characteristics consistently, i.e.,* $\mathsf{P}\left[\hat{b} =_s b_0\right] \to 1$.

The proof of Theorem 2 is given in the appendix. Theorem 2 requires satisfying $\frac{\lambda_n}{\sqrt{n}}\sqrt{J_1} \to 0$, provided that J_1 will not grow much faster than n, which is supported by Condition 2. Further, we impose partial orthogonality along with the irrepresentable condition. These conditions make Theorem 2 going beyond the fixed dimension case.

3.2 Asymptotic analysis under sample selection bias

The presence of sample selection bias poses a challenge to the asymptotic analysis of the variable selection equation. The main challenge of characterising the limiting distribution arises from the asymptotic covariance between $\hat{\beta}$ and $\hat{b}$ because both parameter vectors are high-dimensional and regularised for desired sparsity. The strategy to establish asymptotic analysis under sample selection bias comprises two steps. The first step is to understand the asymptotic normality of the nuisance function estimation error for the sample selection equation (i.e., $\sqrt{N}(\hat{\beta} - \beta)$) and the asymptotic normality of the estimation error of the estimated bias term $\hat{\Gamma} = \hat{\theta}\hat{h}$, where $\hat{h} := h\left(Z^\top \hat{\beta}\right)$ and $\Gamma = \theta h$. This refers to the limiting distribution of the scaled, centered object $\sqrt{N}(\hat{\Gamma} - \Gamma)$, which is the necessary component for establishing the $\sqrt{N}$-consistency and asymptotic normality of our final structural parameter estimator. In the second step, we incorporate the asymptotic variance of $\hat{\beta}$ and asymptotic variance of $\hat{\theta}$ into the asymptotic analysis developed for the variable selection equation that involves the bias correction term. The asymptotic normality in the first step, compared to that in the second step, has an oracle advantage given a larger sample size (full sample N) available in the sample selection equation, namely the estimation of θ does not bring additional contribution to

the asymptotic variance. By contrast, the variable selection equation relies on partially observed samples, provided $n < N$.

For the asymptotic normality of the sample selection bias function, we borrow the asymptotic theorem of Fan and Li (2001) for penalised likelihood estimation where we employ the penalised log-likelihood for the probit lasso model. We denote the penalised log-likelihood $\tilde{\Lambda}(\beta) = \Lambda(\beta) + \sum_{j=1}^{p} pen_{\lambda}(|\beta_j|)$ as the sum of the negative log-probit likelihood $\Lambda(\beta)$ and penality function under regularisation parameter λ. Let $\beta_0 = (\beta_{10}^{\top}, \beta_{20}^{\top})^{\top}$ be a true coefficient with the true nonzero coefficient β_{10} and true zeros captured by β_{20}. Denote the score function $\phi\left(Z_i, \beta_{10}\right) \equiv \partial_{\beta}\tilde{\Lambda}\left(Z_i, \beta_{10}\right)$ and Jacobian matrix $\mathcal{J}_{\beta_{10}} = \partial_{\beta_1} \mathrm{E}\left[\phi\left(Z, \beta_1\right)\right]|_{\beta_1=\beta_{10}}$ to arrive at the influence function $\bar{\phi}\left(Z_i\right) := -\mathcal{J}_{\beta_{10}}^{-1}\phi\left(Z_i, \beta_{10}\right)$. The score function $\phi(Z_i, \beta_{10})$ is evaluated at the true value of the full coefficient vector $\beta_0 = (\beta_{10}^{\top}, \mathbf{0}^{\top})^{\top}$, where $\mathbf{0}$ corresponds to the coefficients of the penalized variables that are assumed to be zero under the sparsity assumption. The L^1 penalty is singular at the origin and does not have continuous second-order derivatives, but can be locally approximated by a quadratic function, see Fan and Li (2001). The resulting score function $\phi(Z_i, \beta)$ is smooth in the neighborhood of β_0. Following their regularity conditions, we obtain the asymptotic normality of β_1 if the irrelevant covariates are known:

$$\sqrt{N}\big(\hat{\beta}_1 - \beta_{10}\big) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \bar{\phi}(Z_i) \to \mathcal{N}(0, \sigma_{\phi}^2), \tag{20}$$

where $\sigma_{\phi}^2 = \mathcal{J}_{\beta_{10}}^{-1}\mathsf{E}\big[\phi(Z, \beta_{10})\phi(Z, \beta_{10})^{\top}\big](\mathcal{J}_{\beta_{10}}^{-1})^{\top}$ is the asymptotic variance of $\hat{\beta}_{10}$. The regularity conditions for $\lambda_N/N \to 0$ lead to a vanishing first-order and second-order derivative of the penalty function. In that case, σ_{ϕ}^2 is asymptotically close to the inverse $\mathbf{I}(\beta_{10})^{-1}$ of the Fisher information matrix.

For the asymptotic normality of the bias correction term, we define a conditional mean function for the bias correction term in a parametric form from (7): $\Gamma(W_i; \theta, \beta) = \mathsf{E}\big[\epsilon_i \big| X_i, Z_i, D_i = 1\big] = \theta h(Z_i\beta)$. Importantly, $\Gamma(W_i; \theta, \beta)$ is differentiable at $\Theta = (\theta, \beta)$ if there exists a linear map (matrix) $\Gamma(\Theta + \ell) : \mathbb{R}^p \to \mathbb{R}^m$ such that $\Gamma(\Theta + \ell) - \Gamma(\Theta) = \Gamma'_{\Theta}(\ell) + \mathbf{o}(\|\ell\|), \quad \ell \to 0$. The derivative map $\ell \to \Gamma'_{\Theta}(\ell)$ is matrix multiplication by the matrix $\Gamma'_{\Theta} = \mathsf{E}\big[\partial_{(\theta', \beta')'}\Gamma\big(W; \theta, \beta\big)\big] = \begin{bmatrix} \Gamma'_{\theta\theta} & \Gamma'_{\theta\beta} \\ \Gamma'_{\beta\theta} & \Gamma'_{\beta\beta} \end{bmatrix}$. Denote the asymptotic covariance matrix for $\Theta = (\theta, \beta)$,

$$\Sigma_{\Theta} = \begin{bmatrix} \sigma_{\psi}^2 & \Omega \\ \Omega & \sigma_{\phi}^2 \end{bmatrix}, \tag{21}$$

where $\Omega = \mathcal{J}_0^{-1}\mathsf{E}_N\big[\psi(W, \theta_0, \eta_0) \otimes \phi(Z, \beta_{10})^{\top}\big](\mathcal{J}_{\beta_{10}}^{-1})^{\top}$ is the outer product between the score column vector $\psi(W, \theta_0, \eta_0)$ and the score row vector $\phi(Z, \beta_{10})^{\top}$, multiplied by $\mathcal{J}_0^{-1}$ and $(\mathcal{J}_{\beta_{10}}^{-1})^{\top}$ as a result of

the delta method. The following proposition gives the asymptotic analysis of the bias error correction.

Proposition 2. *Suppose that the sample selection bias function* $\Gamma(W_i;\theta,\beta)$ *is differentiable at* $\Theta=(\theta,\beta)$ *and* $h_i := h(Z_i\beta)$ *is a twice continuously differentiable function of* β*. If there exists a linear derivative map* Γ'_Θ*, the estimate of* Γ *has asymptotic normality*

$$\sqrt{N}\big(\hat{\Gamma}-\Gamma\big) \xrightarrow{d} \mathcal{N}(0,\Gamma'_\Theta\Sigma_\Theta(\Gamma'_\Theta)^\top), \tag{22}$$

based on the asymptotic covariance defined in (21).

For the asymptotic analysis of the variable selection equation in the second step, we introduce the following conditions.

Condition 6 (Product of kernels). *Define* $\Sigma_{\mathbf{k}}=\frac{1}{n}\mathbf{k}^\top\mathbf{k}$ *the covariance matrix of kernel matrix* $\mathbf{k}\in\mathbb{R}^{n\times JL}$ *that stacks* $\mathbf{k}_i\in\mathbb{R}^{1\times JL}$ *defined after equation (2) for* $i=1,\cdots,n$*. Define a submatrix of* $\mathbf{k}$ *to be* $\mathbf{k}_1\in\mathbb{R}^{n\times J_1L}$ *for* $j\in\mathsf{A}_b$ *and* $|\mathsf{A}_b|=J_1$*. Let* $\Sigma_{\mathbf{k}_1}=\frac{1}{n}\mathbf{k}_1^\top\mathbf{k}_1$ *be a* $J_1L\times J_1L$ *matrix and* J_1 *is diverging in* n*. As* n *increases, we assume* $\Sigma_{\mathbf{k}_1}\to\mathcal{C}$*.*

The product matrix $\Sigma_{\mathbf{k}}\in R^{JL\times JL}$ of kernels is positive definite because of positive definite kernels, implying that $\Sigma_{\mathbf{k}}$ is invertible in spite of a high-dimensional $\mathbf{k}$. The same remark applies to the submatrix $\Sigma_{\mathbf{k}_1}$ by construction.

Condition 7 (Moment condition of kernels). *Bounded moment* $\frac{1}{n}\sum_{i=1}^n\|\mathbf{k}_i\|^2=O_P(1)$

Theorem 3 (Asymptotic normality of characteristics estimate). *Under Conditions 5 – 7 and assumption 1 and* $\lambda_n/\sqrt{n}\to 0$*, with an inclusion of bias correction term* $\hat{\Gamma}:=\hat{\theta}\hat{h}$ *in Proposition 2 and asymptotic properties of sample selection estimators in (20), the adaptive KG-lasso estimates* $\hat{b}=\big(\hat{b}_j\big)_{j=1,\cdots,J}$ *with* $\hat{b}_j:=\hat{b}_j(\hat{\theta})$ *in (19) are consistent and possess the oracle properties*

$$\sqrt{n}\big(\hat{b}-b_0\big) \xrightarrow{d} \mathcal{N}\Big(\mathbf{0},\big(\sigma^2+\Gamma'_\Theta\Sigma_\Theta(\Gamma'_\Theta)^\top\big)\Sigma_{\mathbf{k}_1}^{-1}\Big). \tag{23}$$

Without bias correction, the asymptotic variance of b is inconsistent, which may mislead the estimated significance level. For $\theta\neq 0$, it implies that the null of the sample selection bias test in (18) is rejected. In (23), we obtain an augmented variance induced by $\big(\Gamma'_\Theta\Sigma_\Theta(\Gamma'_\Theta)^\top\big)\Sigma_{\mathbf{k}_1}^{-1}$. The proof of Theorem 3 is provided in the appendix.

4 Empirical results

4.1 Data description

Carbon emission data Following Aswani et al. (2024), Bolton and Kacperczyk (2021, 2023), Sautner et al. (2023) and Zhang (2025), among many others, we employ the *Trucost* database to analyze carbon emissions over a sample period from January 2010 to December 2023. Our sample is limited to US-based firms with common stocks (share codes 10 and 11) traded on the NYSE, AMEX, or NASDAQ exchanges, and return data available in the Center for Research in Security Prices (CRSP) database. *Trucost* provides a comprehensive set of carbon emissions data, widely used by researchers and practitioners, covering scope 1, 2, and 3 emissions as defined by the Greenhouse Gas Protocol.[9] Scope 1 emissions include direct emissions from firm operations, scope 2 accounts for indirect emissions from purchased electricity and other inputs, and scope 3 encompasses other indirect emissions associated with firm supply chain. Additionally, *Trucost* offers metrics on carbon intensity, expressed in equivalent tons of CO_2 (tCO2e) per million dollars of revenue. *Trucost* coverage begins with the fiscal year-end of 2005, and our sample period extends to the fiscal year-end of 2023. In 2016, the coverage was substantially expanded to include small and mid-cap stocks. However, due to limitations in the availability of effective year/month emissions data, practical coverage only starts in May 2009, even for firms with a fiscal year-end of 2005. Thus, we begin our sample period in January 2010. We use fiscal year-end timing and apply a six-month lag to obtain monthly emissions, aligning with standard practices (Fama and French (1993), Bolton and Kacperczyk (2021, 2023)). *Trucost* provides emissions data from a diverse array of sources, including both firm-disclosed and vendor-estimated figures. In total, *Trucost* employs 28 distinct sourcing methods. A substantial fraction of firms consistently receive emissions data derived from a single method throughout the sample period; however, some firms have emissions data compiled from multiple sources. To isolate estimated emissions, we follow Aswani et al. (2024) and retain each of 28 sourcing methods where "estimate" is clearly stated. Busch et al. (2022) report that while emissions data from vendors, including *Trucost*, exhibit near-perfect correlation for disclosed emissions, the correlation for estimated emissions is only around 0.70.

In Figure 1, we report the time-series of total firms covered by *Trucost*. Until 2016, carbon data for around 1000 firms is available with equally disclosing firms and firms for which emissions are estimated. After 2016, the number of firms increased substantially to 2500 and even more than 3000 firms by the end of the sample. Yet, the fraction of firms for which emissions are estimated increased

[9] Further details on protocol standards can be found at https://ghgprotocol.org.

and reached 2/3 by the end of the sample. Overall, a vast majority of firms in the data sample have their emissions estimated.

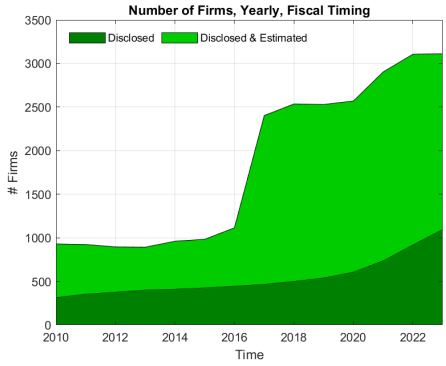


Figure 1: Number of firms over the sample period 2010 to 2023.

Sample expansion has a natural impact on the average firms emissions as illustrated in Figure 2. Scope 1 and 3 are of similar magnitude, even though the latter decreased drastically after the sample expansion in 2016 where average emissions more than halved over the full sample. Even though the impact on average estimated emissions is substantial, it is not less sizeable for disclosed emissions. In the latter case, we observe a steady decrease in average emissions per firm both for scope 1 and 2.

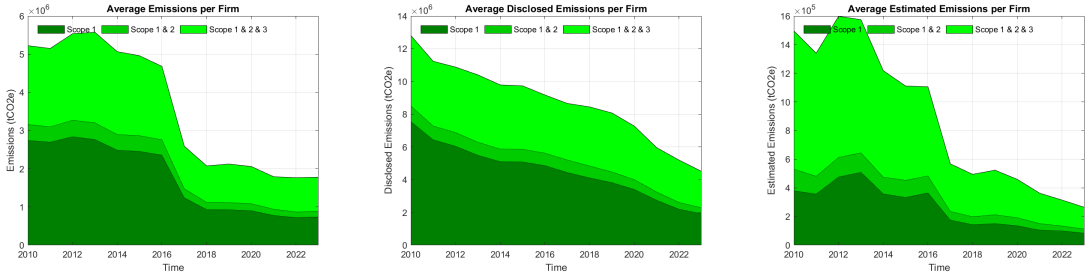


Figure 2: Total Emissions over the sample period 2010 to 2023.

Stock returns and characteristics We obtain monthly stock characteristics from the open-source data provided by Chen and Zimmermann (2022).[10] We begin with a set of 209 characteristics from Chen and Zimmermann (2022) and complement these with monthly stock returns and market capitalization data spanning January 2010 to December 2023 from the CRSP database. We then construct size defined as the logarithm of total equity market value. In our empirical analysis, we retain only those characteristics with at least 25% non-missing data over the sample period to leave

[10] https://www.openassetpricing.com/

us a total of 173 characteristics, and we impute the remaining missing values using mean imputation as advocated by Chen and McCoy (2024).

In Table 1, we report the yearly disclosure rate as well as firm size and age. We standardise the size and age variables at an annual basis. Not surprisingly, the average size of disclosing firms is much larger than that of non-disclosing firms, and disclosing firms are obviously older than non-disclosing firms.

Year	Disclosure rate	Disclosed median	Estimated median	Disclosed std	Estimated std	Disclosed size	Estimated size	Disclosed age	Estimated age
2010	0.305	12.79	10.924	2.987	2.222	0.624	-0.273	0.542	-0.237
2011	0.357	12.546	10.707	2.856	2.157	0.602	-0.335	0.487	-0.270
2012	0.412	12.321	10.645	2.869	2.187	0.527	-0.369	0.374	-0.262
2013	0.438	12.156	10.513	2.847	2.219	0.473	-0.369	0.378	-0.295
2014	0.462	11.93	10.510	2.919	2.195	0.457	-0.393	0.328	-0.282
2015	0.435	11.843	10.453	2.944	2.190	0.527	-0.406	0.382	-0.294
2016	0.459	11.899	10.330	2.905	2.236	0.459	-0.390	0.365	-0.310
2017	0.385	11.759	9.926	2.801	2.318	0.666	-0.417	0.455	-0.284
2018	0.194	11.835	8.799	2.747	2.647	1.171	-0.283	0.808	-0.195
2019	0.203	11.842	8.727	2.831	2.680	1.099	-0.280	0.733	-0.187
2020	0.223	11.687	8.662	2.883	2.674	0.986	-0.283	0.686	-0.197
2021	0.245	11.598	8.522	2.859	2.668	0.903	-0.294	0.664	-0.216
2022	0.267	11.360	7.941	2.888	2.664	0.890	-0.324	0.619	-0.226
2023	0.326	10.925	7.798	3.006	2.621	0.799	-0.387	0.502	-0.243

Table 1: Disclosure Rate and Firm Size

As already discussed, the disclosure rate is particularly low (around 25%) after the sample expansion in 2016. Interestingly, over the whole sample period, disclosing firms are always large and established firms. Yet, the correlation between emissions and size are stronger in estimated emissions than in disclosed emissions as shown in Figure OA.1. The high correlation with size raises serious multicollinearity issues when measuring the carbon premium as extensively discussed by Aswani et al. (2024) and Zhang (2025).

4.2 Sample selection results

We begin with the sample selection equation (5) and obtain the estimates of β by lasso probit in (9) that allows us to handle high-dimensional Z_i and undertake variable selection in the propensity score parametrised with $Z_i\beta$. The parameter in the lasso penalty is chosen by 5-fold cross-validation. We present the sample selection results in Figure 3 for the non-zero $\hat{\beta}_j$ (active characteristics).[11]

[11]The constellation of $\hat{\beta}$ estimates can be found in Figure OA.2.

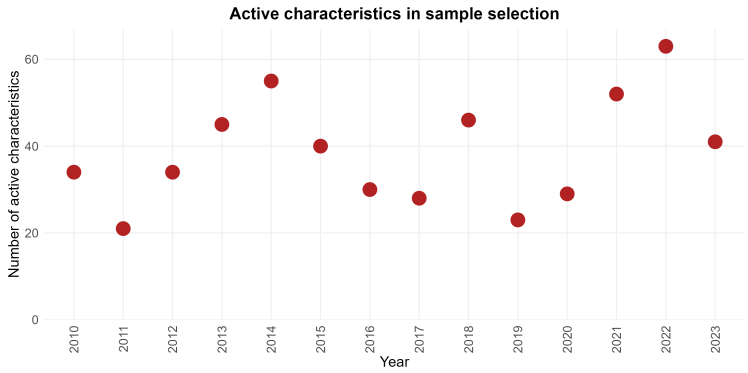


Figure 3: Number of active characteristics in sample selection

While the number of characteristics needed to predict disclosure fluctuates over time, it increases substantially across the sample period and roughly doubles between the beginning and the end of the sample. This increase partly reflects the progressive inclusion of smaller and less established firms following the expansion of *Trucost* coverage, for which disclosure decisions are more sensitive to firm-specific constraints, informational frictions, and cost considerations.

We find that three characteristics repeatedly play a central role in sample selection: "Size" (measured by the logarithm of market capitalization), "Age" (measured by the number of months the firm is present in CRSP), and two-month lagged trading volume in dollars ("DolVol"). These variables capture three fundamental economic dimensions that are known to shape firms' disclosure incentives: scale, information quality, and market liquidity. Larger firms face lower fixed costs of disclosure and greater regulatory and investor scrutiny; older firms typically operate in more transparent information environments with more stable reporting practices; and more liquid firms are more closely monitored by market participants, increasing the benefits of disclosure relative to opacity. Consistent with these mechanisms, firms with higher market capitalization, longer market presence, and greater trading activity are more likely to voluntarily disclose emissions. This pattern aligns with existing evidence in the disclosure and sustainable finance literature, including Flammer et al. (2021), Gibbons (2024), Krueger et al. (2024), and Gehricke et al. (2025).

The relatively higher disclosure propensity among large and liquid firms can be rationalized by standard cost–benefit considerations. On the cost side, such firms are better equipped to measure and aggregate emissions - either directly or through internal reporting systems - making disclosure less costly. On the benefit side, disclosure allows them to mitigate regulatory risk, respond to investor demand for transparency, and credibly signal compliance with environmental norms. In contrast, smaller, younger, or less liquid firms face higher relative disclosure costs and weaker market discipline, making non-disclosure a more attractive option.

We also find that profitability, as measured by analyst earnings per share ("FEPS"), growth opportunities, proxied by sales growth and relative ranking among peers ("MeanRankRevGrowth"), and pension funding ("FR"), a proxy for long-term commitments to employees, are important predictors of disclosure, appearing in 12 out of 14 years. These variables are commonly associated with firm quality, financial slack, and long-term orientation, all of which reduce the expected costs of disclosure and the perceived risk of revealing environmental information. Only 31 out of the 173 characteristics are never selected over the sample period, underscoring the importance of a high-dimensional framework to capture the multifaceted determinants of voluntary disclosure. Importantly, while these firm-quality and stakeholder-related characteristics are relevant for the disclosure decision, they need not directly determine carbon emissions themselves, and they are indeed not selected in the variable selection equation, as discussed in the next section.

Coefficient of sample selection bias term To correct for sample selection bias from the sample selection equation, we proceed with estimating θ for each calendar year using Neyman orthogonal score defined in (15). It turns out that the M-estimator θ satisfies the moment condition specified in (16). We deploy the DML estimation strategy detailed in Algorithm 1 where we adopt a 5-fold cross-fitting. Each run of Algorithm 1 renders one estimate of θ, and a thirty-run gives us the distribution of θ in Figure 4. The DML estimation technique relies on subsamples obtained by randomly partitioning the sample: an auxiliary sample for estimating the nuisance functions and a main sample for estimating the parameter of interest. To incorporate the impact of sample splitting, the mean of θ distribution is suggested by Chernozhukov et al. (2018) in their Definition 3.3.

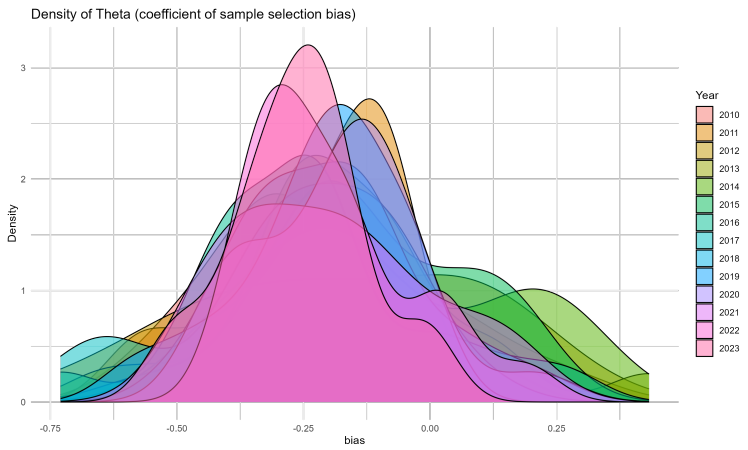


Figure 4: Distribution of θ across year 2010-2023 for Scope 1

It appears that before 2017, the uncertainty induced by sample splitting is relatively higher than in recent years, as clear from the dispersion of the distribution. It reflects a relatively smaller

sample size and a lower disclosure rate in the early period. The location of distributions indicates a negative value of $\hat{\theta}$, implying a negative correlation between the error process in (4) and that of (5). Unobserved factors increasing selection probability are negatively correlated with unobserved factors affecting the outcome, which is emission level in log. As an example, the firms possessing unobserved green characteristics are inclined to disclose, but also more likely to yield a reduced scale of emission output. In Table 2, we showcase the statistical significance of $\hat{\theta}$ and its variance estimator (see Theorem 3.2 of Chernozhukov et al. (2018)) for scope 1 emissions.

	$\hat{\theta}$	σ_ψ	$\frac{\hat{\theta}}{\sigma_\psi}$	p-value $\frac{\hat{\theta}}{\sigma_\psi}$	s_n	p-value s_n
2010	-0.208	0.028	-7.432	0.000	5.034	0.025
2011	-0.211	0.034	-6.244	0.000	9.186	0.002
2012	-0.179	0.023	-7.643	0.000	7.648	0.006
2013	-0.166	0.021	-7.760	0.000	7.794	0.005
2014	-0.089	0.022	-4.098	0.000	10.624	0.001
2015	-0.125	0.023	-5.548	0.000	8.508	0.004
2016	-0.243	0.019	-12.615	0.000	8.341	0.004
2017	-0.257	0.018	-14.034	0.000	7.974	0.005
2018	-0.196	0.022	-8.775	0.000	8.071	0.004
2019	-0.219	0.012	-18.471	0.000	5.467	0.019
2020	-0.180	0.011	-15.669	0.000	5.805	0.016
2021	-0.198	0.016	-12.153	0.000	7.163	0.007
2022	-0.197	0.017	-11.673	0.000	8.047	0.005
2023	-0.259	0.019	-13.635	0.000	10.029	0.002

Table 2: Significance of $\hat{\theta}$ and score test for Scope 1

Obviously, all the estimates of $\hat{\theta}$ are statistically significant at 1% level, indicating that the sample selection bias is statistically present. The same pattern occurs when looking at scope 2 (Table OA.1) or scope 3 emissions (Table OA.2). We corroborate that with the hypothesis test of the sample selection bias in (18) where we test the null hypothesis $H_0 : \theta = 0$ against the alternative hypothesis $H_1 : \theta \neq 0$. The critical value at 95% confidence level is obtained from the chi-square distribution with one degree of freedom, which is 3.841. The null hypothesis is rejected at 5% level for all calendar years. To assess the impact of regularisation bias, we also report the estimates of $\hat{\theta}$ obtained from the model without mitigating that potential effect. We display the estimates obtained without sample splitting and K-fold cross-fitting in Table OA.3. We find that $\hat{\theta}$ is not always statistically significant across years. However, our score test remains powerful and rejects the null hypothesis of no sample selection bias for all years. The estimates of $\hat{\theta}$ are still negative, but the magnitude is smaller than that obtained from the DML approach correcting for the regularisation bias. It indicates that such a bias is not negligible and should be accounted for.

4.3 Variable selection results

We define two models: **M1**: variable selection *with* sample selection bias correction; **M2**: variable selection *without* sample selection bias correction, and compare the variable selection results. The nonzero coefficients $\hat{b} \neq 0$ per portfolio and per year estimated by **M1** and **M2** for Scope 1 are displayed in Figure 5.

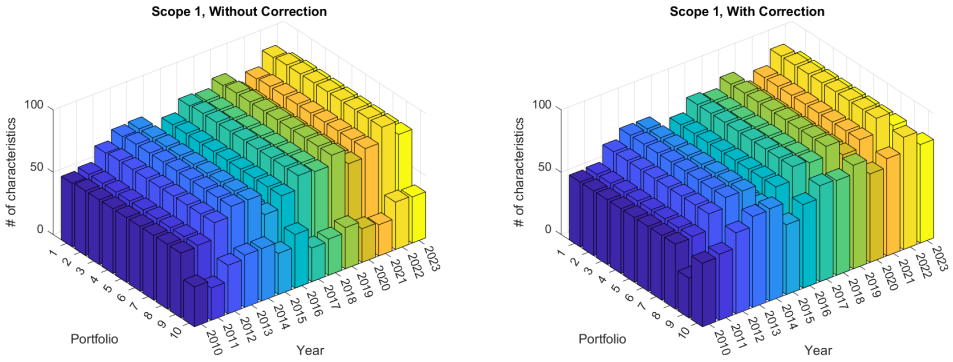


Figure 5: Active Characteristics Portfolios in Variable Selection, Scope 1

As for sample selection, the relevance of firm characteristics increases in variable selection over time, partially driven by the greater heterogeneity introduced by the sample expansion in 2016. As smaller and more diverse firms enter the dataset, emissions become increasingly heterogeneous across firms, requiring a richer set of characteristics to capture underlying production technologies and emission processes. This pattern is consistent across portfolios and holds both with and without the application of sample selection bias correction.

A critical role of the bias correction is to restore the relevance of extreme characteristic-sorted portfolios, in particular the top decile. While these portfolios exhibit limited explanatory power without correction, they become as relevant as other portfolios once selection bias is accounted for. This result is economically important. Extreme portfolios (top and bottom deciles) constitute the fundamental building blocks of long–short factors widely used in empirical asset pricing and corporate finance (Fama and French; 1993), as well as in the measurement of the carbon premium (Pástor et al.; 2022; Avramov et al.; 2025). Muting either the long or the short leg of such factors through selection-induced attenuation is therefore problematic and can lead to biased inference.

An intuitive economic explanation for this result is that firms located in the extreme tails of characteristic distributions are precisely those for which disclosure incentives are most distorted. Firms in the top or bottom deciles typically exhibit unusually high or low levels of size, investment intensity, growth opportunities, or financial constraints. For these firms, the costs and benefits of disclosure

are highly asymmetric, making voluntary disclosure particularly endogenous. In the absence of correction, extreme firms are therefore disproportionately underrepresented among disclosers, leading to a truncation of the effective support of characteristics in the observed sample. This truncation mechanically attenuates the explanatory power of extreme portfolios in the outcome equation.

Once selection bias is corrected, the model effectively reconstructs the counterfactual contribution of firms in the tails of the characteristic distribution, allowing extreme portfolios to recover their economic relevance. This mechanism is consistent with the interpretation of selection bias as a form of endogenous censoring: without correction, observed emissions predominantly reflect firms with intermediate characteristics, while firms at the extremes - which often drive cross-sectional variation in emissions - are partially hidden by non-disclosure. Thus correcting for selection restores the role of extreme portfolios as carriers of economically meaningful variation, in line with their central role in long–short factor construction and carbon premium measurement.

The most striking result is the absence of any role for portfolios sorted on firm size, despite size being pivotal in sample selection as revealed in the previous section. While disclosure decisions are strongly driven by size - reflecting fixed disclosure costs, regulatory scrutiny, and information environments - size does not directly govern emissions conditional on participation once a rich set of firm characteristics is taken into account. Although this overlap could raise identification concerns, the adaptive weighting scheme effectively imposes exclusion restrictions by penalizing characteristics that primarily affect participation rather than emissions. After bias correction, only one additional characteristic remains inactive, namely the six-month zero-trade indicator, further underscoring the necessity of a high-dimensional analysis. In the uncorrected specification, however, thirteen characteristics are inactive.[12] Overall, less than 10% of the portfolios comprise the list of mismatched characteristics selected in **M1** and **M2**. The number of mismatches increases from 86 at the beginning of the period to 146 in 2022, to be compared to 1,730 portfolios each year.

Taken together, the contrast between the uncorrected and selection-corrected specifications highlights a clear economic distinction between characteristics that primarily govern disclosure decisions and those that shape emissions conditional on disclosure. Variables related to market liquidity, trading frictions, firm visibility, and information environments - such as size, trading volume, zero-trade indicators, or return distribution moments - play a central role in participation decisions by prox-

[12]Without bias correction, inactive characteristics are: Beta adjusted to idiosyncratic volatility ("BetaFP"), Coskewness with market ("CoskewACX"), Firms omitting to pay dividend ("DivOmit"), Trading volume ("DolVol"), Exchange switching ("ExchSwitch"), Amihud's illiquidity measure ("Illiquidity"), Long term momentum ("MomSeason11YrPlus"), Delayed stock to market reaction ("PriceDelayTstat"), IPO without RD ("RDIPO"), Skewness of returns ("ReturnSkew"), Size ("Size"), Average over 12 months of number of days without trade ("zerotrade12M") and Average over 6 months of number of days without trade ("zerotrade6M"). With bias correction, only $zerotrade6M$ and $Size$ remain inactive.

ying for monitoring intensity, investor attention, and fixed disclosure costs. When selection bias is ignored, endogenous disclosure mechanically attenuates the contribution of characteristics that operate through real production and financing channels, leading to their apparent irrelevance in the outcome equation. Once selection bias is accounted for, the model recovers characteristics linked to firms' dynamic adjustment margins, including investment, innovation, external financing, and capital structure choices. These variables directly affect production scale, capital intensity, and technological upgrading, and therefore emissions conditional on participation. Conversely, characteristics that remain inactive even after correction do not represent structural drivers of emissions once real activities are properly controlled for. This separation between participation-related and outcome-related characteristics is consistent with the identifying structure of the model and underscores the importance of correcting for endogenous disclosure when inferring the economic determinants of corporate emissions.

The key characteristics for Scope 1, namely the ones selected 85% of the time over the full sample period, relate to future growth opportunities (R&D, profitability, investment) and capital structure (external financing, leverage and payout indicator).[13] Figure OA.3 presents the estimated coefficients of the selected characteristics in **M1** and **M2**. The findings for Scope 2 and 3 are qualitatively similar, particularly concerning the limited importance of the size characteristic. As shown in Figure OA.4, Scope 3 begins with a relatively large number of relevant firm characteristics, which declines slightly toward the end of the sample period. This is consistent with the nature of Scope 3 emissions, which account for a significant portion of total firm emissions and reflect value chain activities, making them more complex to measure and therefore requiring a broader set of explanatory variables. While not all characteristics remain consistently relevant over time, the number of persistently selected key characteristics is 10 for Scope 1, 5 for Scope 2, and 10 for Scope 3. Two variables - Debt Issuance and R&D - are common across all scopes, highlighting the central role of financing and technological investment in shaping firms' emissions profiles. The variation in selected characteristics across scopes indicates that each captures distinct economic dimensions of firm emissions. We have also checked that including the characteristic "Industry Sector" does not alter the results. Indeed, that additional characteristic is never selected for the variable selection equation and therefore plays no role in the imputation process. It is only selected for the sample selection equation (2 out of 14 years), but without material impact on the empirical results.

[13] Adjusting for selection bias introduces the top deciles of the following variables into the selection: Credit rating downgrade ("CredRatDG"), Earning forecasts ("FEPS"), Unexplained book-to-market ratio ("Frontier"), IPO occurring ("IndIPO"), Deflated investment growth ("InvGrowth"), Customer oriented industries momentum ("iomom_cust"), Supplier oriented industries momentum ("iomom_supp"), Long term debt leverage ("NetDebtFinance"), Option relative to equity volume ("OptionVolume2"), IPO without R&D ("RDIPO"), Growth in number of shares ("ShareIss5Y"), New stocks ("Spinoff") and Trading volume over total market capitalization ("VolMkt").

4.4 Emission prediction performance

We show the emission prediction performance between **M1** and **M2** over time in Table 3. For each calendar year, we randomly split the disclosed sample into training and testing sets (90% vs. 10%). We train **M1** and **M2** using the training set and predict the emission outcomes for the testing set. We report the MSE of the testing set for each calendar year and the relative error as the ratio of MSE under **M1** and **M2** .

Year	MSE of **M1**	MSE of **M2**	Relative error
2010	2.553	4.235	0.603
2011	3.117	3.399	0.917
2012	2.742	3.259	0.841
2013	3.498	4.109	0.851
2014	3.298	4.185	0.788
2015	1.898	2.739	0.693
2016	2.294	3.319	0.691
2017	2.262	2.726	0.830
2018	3.174	3.946	0.804
2019	2.939	3.186	0.923
2020	4.275	5.215	0.820
2021	2.362	3.585	0.659
2022	3.109	3.893	0.799
2023	2.808	3.593	0.782

Table 3: Prediction performance (Scope 1)

M1 demonstrates superior predictive performance, with relative MSE ratios ranging from 0.603 to 0.923. Prediction results for the testing set of disclosure samples under **M1** and **M2** are shown in panels (a) and (b) of Figure OA.5. Compared to **M2**, **M1** displays a wider interquartile range, likely due to the incorporation of the bias correction term. Owing to negative values of $\hat{\theta}$, the median predicted values under **M1** are lower than those under **M2**. This suggests that omitting the correction for sample selection bias leads to an overestimation of disclosed emissions. The same pattern holds for scope 2 and 3 emissions, as shown in Tables OA.5 and OA.6. Interestingly, the MSE across all years is lower for scope 3 emissions than for scope 1 and 2, under both **M1** and **M2**. This is encouraging, given that scope 3 is generally the hardest to measure and accounts for a significant share of total emissions. At the same time, scope 3 emissions are also the category where the sample selection bias correction, albeit significant from Table OA.2, has the least impact in terms of relative prediction performance (relative error close to 0.9 for the majority of years in Table OA.6). One possible explanation for this pattern is that, even when firms disclose scope 3 emissions, they typically rely on estimation procedures based on characteristics of upstream and downstream

partners. As a result, disclosed scope 3 emissions - despite being estimated by firms - are broadly consistent with what could be inferred using publicly available information.

One of the advantages in the `SS-VS` model is to predict the undisclosed emission given the estimated propensity of firm disclosure decision and the resulting sample selection bias. To be more explicit, the undisclosed emissions can be estimated under **M1** via

$$\hat{Y}_i = \hat{\mathsf{E}}\big[Y_i\big|X_i, Z_i, D_i = 0\big] = \mathbf{k}_i\hat{b} + \hat{\mathsf{E}}\big[\epsilon_i\big|X_i, Z_i, D_i = 0\big] = \mathbf{k}_i\hat{b} - \hat{\theta}h(Z_i\hat{\beta}). \tag{24}$$

Panels (c) and (d) of Figure OA.5 display the box plot of prediction for the undisclosed samples. Panel (c) shows the results of **M2** in comparison with the results of **M1** in (d). We observe that the median and interquartile range in (c) and (d) are generally smaller than those in (a) and (b) using the disclosed samples. It is not surprising because the disclosed samples have bigger firm sizes, and those firms tend to produce a larger scale of emission compared to small firms, as evident in Table 1. We also compare (c) and (d) with the *Trucost* estimates. It appears that the *Trucost* estimates look indifferent between 2010-2016 (same median and interquartile range for these years), along with another block of estimates with high persistence during 2017-2023 in terms of interquartile range. More importantly, the *Trucost* estimates potentially yield many outliers in the left-tailed distribution.

The idea in the estimation of the undisclosed emissions through (24) is similar to the use of a parametric model to describe the probabilistic behaviour of the censored or truncated (unobserved) parts in duration data, such as unemployment spells. Here, the term $\mathsf{E}\big[\epsilon_i\big|X_i, Z_i, D_i = 0\big]$ is not identified nonparametrically, and we cannot extrapolate unobserved emissions from observed emissions without additional structural assumptions; see Manski (1989), Bia et al. (2024), and Huber and Melly (2015) for further discussion. Hence, to mitigate the potential risk of misspecification inherent to the normality assumption, we also opt for a flexible control function approach similar to Newey (2009) as a robustness check. Rather than imposing a specific distribution on the latent errors, we approximate the unknown selection process using a power series expansion in the IMR. Specifically, we define the conditional expectation of the error term for the disclosing sample as $\mathsf{E}[\epsilon_{it} \mid X_{it}, Z_{it}, D_{it} = 1] = \sum_{m=1}^{M} \rho_m h_i^m$, where M is the order of the polynomial expansion. When $M = 1$, the outcome equation boils down to (8) with $\rho_1 = \theta$. If we let M grow with n as in Assumption 4.5 of Newey (2009), we get a semi-nonparametric estimation setting. Here, we prefer to use M fixed to allow us to keep a high-dimensional parametric linear regression as in (8) and use the theory developed in the previous sections. How can we address the counterfactual estimation, namely the challenge of estimating the outcome (carbon) for the non-selected group (counterfactual), under that flexible modeling? By the law of total expectation, the unconditional mean of the error ϵ_i must satisfy: $\mathsf{E}[\epsilon_i] = \mathbb{P}[D_i = 1|Z_i]\mathsf{E}[\epsilon_i|X_i, Z_i, D_i = 1] + \mathbb{P}[D_i = 0|Z_i]\mathsf{E}[\epsilon_i|X_i, Z_i, D_i = 0] = 0$, and the selec-

tion bias for the non-selected group ($D_i = 0$) is $\mathsf{E}[\epsilon_i|X_i, Z_i, D_i = 0] = -\frac{\mathbb{P}[D_i=1|Z_i]}{\mathbb{P}[D_i=0|Z_i]}\mathsf{E}[\epsilon_i|X_i, Z_i, D_i = 1]$, where $\frac{\mathbb{P}[D_i=1|Z_i]}{\mathbb{P}[D_i=0|Z_i]}$ corresponds to the so-called "odds of selection". It means that we can recover the bias of the nondisclosing sample ($D_i = 0$) by scaling the bias of the disclosing sample ($D_i = 1$) with the probability that the firm should have been in the sample but was not. Let $\hat{P}_i = \Phi(Z_i\hat{\beta})$ be the estimated propensity score. The estimated counterfactual mean (24) becomes simply:

$$\hat{\mathsf{E}}[Y_i|X_i, Z_i, D_i = 0] = \mathbf{k}_i\hat{b} - \left(\frac{\hat{P}_i}{1-\hat{P}_i}\right)\sum_{m=1}^{M}\hat{\rho}_m h(-Z_i\hat{\beta})^m, \tag{25}$$

under the series expansion. Empirically, we find that the estimators of ρ_1 and ρ_2 are both statistically significant and negative, whereas higher-order terms do not lead to additional explanatory power. This shows that not only the slope but also the curvature of the bias matters, and both collectively determine the selection process. This confirms our hypothesis of "strategic tail-hiding": the worse a firm true carbon performance, the more aggressively it leverages non-disclosure, leading to further increase in selection bias.

4.5 Pecuniary implications

We assess the pecuniary impact of sample selection bias on the estimation of firm carbon emissions by computing the median difference between the predicted emissions derived from **M1** and **M2** for each firm. This difference is then aggregated across the subset of firms that do not disclose emissions. We report the term $\hat{\theta}h(Z_i\hat{\beta})$ expressed in units of tCO2e, which quantifies the aggregated bias (underestimation) in estimated emissions for non-reporting firms. We also report the difference between our bias selection corrected estimation and *Trucost* estimated emissions.

One relates to the changes in data coverage over our sample period. As more small and medium-sized firms entered the dataset, the median level of absolute carbon emissions mechanically declined over time. To account for this, we present the results in relative terms in Figure 6, scaling the median underestimation by the median emissions of firms that disclosed their carbon emissions. Absolute figures are reported separately in Figure OA.6 and detailed figures in Tables OA.7 - OA.9.

Two notable patterns emerge. First, while underestimation is substantial at the beginning of the sample period (2010), the gap narrows considerably over time for scope 1 and 2 emissions. This convergence is less pronounced for scope 3, once again highlighting the challenges of assessing value chain related emissions. Average underestimation in scope 1, 2, and 3 was 9.28 %, 9.04 %, and 5.01 %. It steadily decreased for scope 1 and 2 (from 13.21 % to 3.84 % and from 9.37 % to 5.30 %),

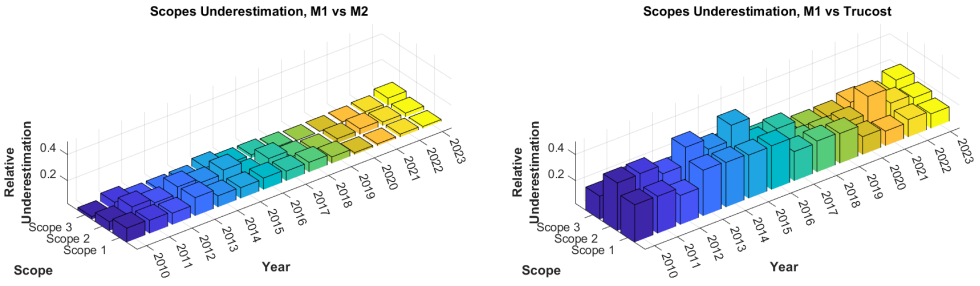


Figure 6: Relative Underestimation in Scope Emissions

but remained fairly stable for scope 3, with even a slight increase toward the end of the sample period. Once again, scope 3 appears to behave differently from the other scopes. To price the impact of selection bias, we also examine absolute values, which are reported in Figure OA.6. The median underestimation for scope 1 falls sharply from 41,977 tCO2e in 2010 to 1,947 tCO2e in 2023 - a 95 % decline. For scope 2, the decrease is less steep in absolute terms (from 27,338 tCO2e to 3,080 tCO2e), but still economically significant (–89 %). For scope 3, the reduction is more modest, with the median underestimation falling by roughly half (from 70,615 tCO2e to 38,719 tCO2e). The total aggregate underestimation over the full period is substantial: 167.6 million tCO2e for scope 1, 197.6 million tCO2e for scope 2, and 676 million tCO2e for scope 3 - adding up to nearly 1 GtCO2e. As we discuss below, this large-scale underestimation and the related selection bias likely carry significant pecuniary implications.

A second noteworthy pattern is that the gap in emissions estimated by *Trucost* closely mirrors - at least qualitatively - the estimates produced by our **M1** and **M2** models when using absolute values. However, when examining relative values, the picture diverges sharply between *Trucost* and **M1**. Average relative underestimation is 26.30 % for scope 1, 30.73 % for scope 2, and 18.33 % for scope 3. While the exact estimation method used by the vendor is unknown, these figures suggest that additional information may be incorporated beyond publicly available firm characteristics. The dynamic pattern for scopes 1 and 2 is relatively similar between *Trucost* and **M1**: underestimation declines from 31.13 % to 13.50 % for scope 1, and from 39.14 % to 17 % for scope 2. In contrast, the evolution for scope 3 is less monotonic, moving from 20.24 % to 21.86 %, with intermediate values as low as 12.19 %. Absolute values further confirm this picture. In 2010, the median *Trucost* underestimation for scope 1 was 98,928 tCO2e, compared to 41,977 tCO2e for **M2**. For scope 2, the figures were 114,161 tCO2e for *Trucost* versus 27,338 tCO2e for **M2**; and for scope 3, 315,008 tCO2e versus 70,615 tCO2e. Although all figures declined markedly over time, *Trucost* underestimation

remained larger than that of **M2** by the end of the sample: 6,840 tCO2e vs. 1,947 tCO2e for scope 1; 9,871 tCO2e vs. 3,080 tCO2e for scope 2; 105,642 tCO2e vs. 38,719 tCO2e for scope 3. These trends suggest a meaningful improvement in data quality over time. Although it remains unclear whether *Trucost* adjusts for selection bias, the narrowing gap between their estimates and ours - particularly for scope 1 and 2 - indicates a clear convergence. Notably, the alignment between *Trucost* estimates and those produced by our models implies that such emissions data can be closely replicated using high-dimensional public information and the estimation procedures developed in this paper. Given the sophistication of our methodology, *Trucost* effort to make these estimates readily available to investors provides significant added value. Although underestimation declines over time, it remains economically meaningful. It is therefore important to quantify the associated pecuniary cost. To compute this monetary gap, we apply an explicit carbon price of €2.35 per tCO2e, sourced from the US Emissions Trading System (ETS) and carbon tax data reported by the OECD.[14] Using the average 2023 euro-to-dollar exchange rate of 0.924, this corresponds to a carbon price of $2.54 per tCO2e.[15] Our findings are reported in Tables OA.7 - OA.9.

The pecuniary impact of emissions underestimation - and the resulting shortfall in carbon tax revenues - is economically significant. For scope 1, the total cost over the sample period amounts to $425.8 million, declining from $62.3 million in 2010 to $9.3 million in 2023. The cumulative figures for scope 2 and 3 are $501.9 million and $1.72 billion, yielding an aggregate shortfall of approximately $2.65 billion. The corresponding figure based on *Trucost* estimates is even larger, reaching nearly $9.5 billion. These magnitudes underscore the economic relevance of underreporting and support our argument regarding green silence - that is, the adverse selection problem posed by non-reporting firms in the carbon disclosure landscape. This issue warrants careful attention from regulators and policymakers seeking to reduce the social cost of carbon emissions. For illustrative purposes, we translate the unreported emissions gap into an equivalent number of transatlantic flights operated by an Airbus A330 on the London (LHR) to New York (JFK) route. A single one-way flight on this route emits approximately 150 tCO2. Framing the emissions gap in this way - equivalent to several thousand flights per year - helps convey the magnitude of underreporting in tangible terms. Over the full sample period, the underestimation due to selection bias corresponds to more than 6 million such flights. Likewise, to assess the impact of regularisation bias, we also quantify the pecuniary impact of sample selection bias on the estimation of firm carbon emissions using $\hat{\theta}$ estimates from Table OA.3. In Table OA.4 , the underestimation of emissions for non-reporting firms is not as sizeable as observed for the DML approach, accounting for regularisation bias in Table OA.7. The pecuniary impact linked to underreported emissions and the associated shortage of carbon tax revenue is relatively lower. Such

[14] https://www.oecd.org/content/dam/oecd/en/topics/policy-sub-issues/carbon-pricing-and-energy-taxes/carbon-pricing-united-states.pdf

[15] https://www.irs.gov/individuals/international-taxpayers/yearly-average-currency-exchange-rates

an evidence documents that neither the sample selection bias nor the regularisation bias should be neglected in the estimation of firm carbon emissions.

Finally, the social cost of capital includes broader considerations and is usually different from the carbon price. It reflects the current value of future damages generated by future carbon emissions. Given the multiplicity of parameters used to compute this social cost, it can lean on a wide range. Van den Bremer and Van der Ploeg (2021) recently provide a risk-adjusted measure of the social cost of capital, from \$6.6 per tCO2e (market-based estimate) to \$66.3 per tCO2e (ethics-based estimate). On the other hand, Pastor et al. (2025) advocate use of the social cost of carbon provided by the US EPA agency (EPA (2023)), which is also in a wide range. A meta-analysis of existing estimates of this cost by Tol (2023) provides a range from \$9 to \$525 per tCO2e. Given that total underestimation is close to 1 GtCO2e, this implies that underestimating carbon emissions is tantamount to underestimating the social cost of carbon by at least \$9 billion - and possibly by as much as \$525 billion for the period 2010-2023.

One final consideration concerns the role of firm size in estimating emissions. As shown above, while size plays a central role in the sample selection process, it does not appear in the variable selection for imputation. This is notable given that prior literature has emphasized the importance of size in the estimation procedures used by data vendors. To explore this further, we repeated our analysis using only firm size or revenue to impute missing emissions data. The results, reported in Table OA.10 for scope 1 and illustrated in Figure 7, highlight the implications of relying on such naïve imputations.

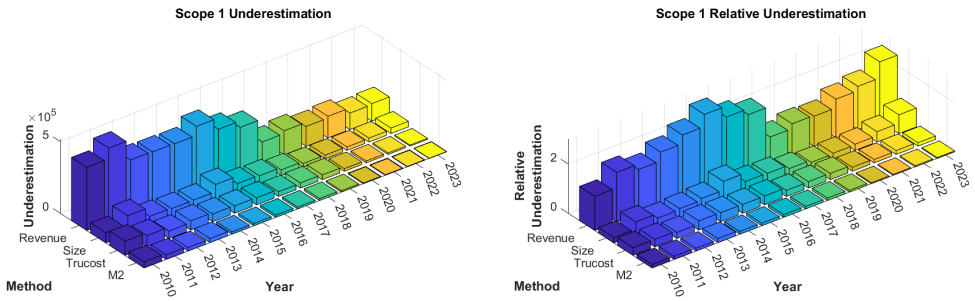


Figure 7: Underestimation in Scope 1 Emissions by Naive imputation.

While the total underestimation under the **M2** model amounts to 167,622,445 tCO2e, this figure rises sharply when using naïve imputation methods - reaching 675,978,583 tCO2e when based solely on firm size, and an even more pronounced 3,250,987,006 tCO2e when using revenue. The associated pecuniary costs increase proportionally. In relative terms, revenue stands as the worst univariate

screening characteristic for carbon emissions imputation. This evidence provides strong support for the use of high-dimensional approaches to correct for selection bias in carbon emissions estimation. Eventually, let us mention that our robustness check based on a series expansion yields slightly milder underestimation numbers than Table OA.7 due to the additional odd selection $\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right)$ term in (25), even though the series expansion inflates the correction size; see Table OA.11 for Scope 1. We also observe akin results for Scope 2 and 3.

4.6 Implications for the carbon premium

Our finding that carbon data vendors under-estimate the emissions of non-disclosers has first-order implications for tests of the existence of a carbon premium. If investors care about firm carbon footprints, prices should reflect this through the cost-of-capital channel: lower expected returns (cheaper capital) for greener firms and higher expected returns (dearer capital) for heavier emitters. Detecting such a premium hinges on correctly classifying firms into high versus low emitters. Downward-biased imputations for non-disclosers mechanically mis-sort some high emitters into cleaner bins, compress cross-sectional dispersion in emissions, and thus may attenuate estimates of the carbon premium (and, in some designs, can even flip its sign).[16] To quantify these consequences, we compare firm classifications and the resulting premium across five constructions: our selection bias uncorrected and corrected estimators, the vendor series and two additional naïve benchmarks, size-scaled and revenue-scaled proxies, that impute non-disclosers from disclosers. This design isolates how measurement quality drives inferences about the carbon premium.

To assess how measurement choices translate into firm classifications, we take the five alternative emissions estimators and, at a monthly frequency, compute the pairwise correlations between the firm-level emissions they imply; we then average these correlations over time and report the resulting time-averaged correlation matrix in Table 4. Beyond correlations in levels, we also report an agreement metric based on the Jaccard index.[17] Using the selection-corrected estimator as our benchmark, we report the Jaccard index between this benchmark and each of the four other methods within emissions-sorted buckets from the bottom 10% (low emitters) through the bottom 30% and 50% and up to the top 10% (high emitters). We focus on Scope 1 and include the corresponding Tables for Scope 2 and 3 in the Appendix.[18]

[16] See contributions by Aswani et al. (2024), Bolton and Kacperczyk (2021, 2023), Zhang (2025).

[17] For two sets A and B, the Jaccard index is $J(A,B) = \frac{|A\cap B|}{|A\cup B|} \in [0,1]$. In our application, A and B denote the sets of firms classified as "low" (or "high") emitters by two methods at a given cutoff; $J = 1$ indicates identical classifications, $J = 0$ indicates disjoint classifications.

[18] See Tables OA.12 - OA.14 (Scope 2) and OA.15 - OA.17 (Scope 3).

	Rank Correlation						Jaccard Index			
	M1	**M2**	**Vendor**	**Size**	**Revenue**		**M2**	**Vendor**	**Size**	**Revenue**
M1	1.00	0.99	0.72	0.55	0.71	Bottom 10%	0.74	0.17	0.22	0.23
M2	0.99	1.00	0.74	0.54	0.70	Bottom 30%	0.84	0.45	0.28	0.46
Vendor	0.72	0.74	1.00	0.42	0.58	Bottom 50%	0.90	0.61	0.47	0.57
Size	0.55	0.54	0.42	1.00	0.62	Top 50%	0.90	0.64	0.47	0.56
Revenue	0.71	0.70	0.58	0.62	1.00	Top 30%	0.87	0.57	0.44	0.55
						Top 10%	0.91	0.71	0.71	0.69

Table 4: Agreement Across Carbon Emissions Estimation Methods.

The correlation matrix shows very high agreement between our two high-dimensional, characteristics-based imputations, the selection-bias corrected estimator (M1) and the version that does not correct for this bias (M2). The selection-bias correction primarily shifts levels rather than reorders firms. By contrast, correlations with the vendor series and with naïve size- and revenue-scaled imputations are much lower. In particular, the size proxy correlates only 0.55 with M1 and 0.54 with M2, underscoring how far naïve scaling is from our double–machine–learning approach; the vendor series is also distant from both our estimators and from the naïve proxies.

To localize disagreement that matters for carbon premium measurement, the Jaccard-index of M1 vs. M2 yields Jaccard values above 0.8 for most buckets with the largest disagreement in the bottom 10% bucket. This indicates that modeling selection is essential for identifying the cleanest firms, which are the focus of many screening strategies. Relative to M1, agreement with alternative methods is markedly weaker in the bottom decile: Jaccard is 0.17 for the vendor, 0.22 for Size, and 0.23 for Revenue. Differences persist among high emitters but are far less pronounced. Overall, measurement choice is most consequential precisely where carbon-premium tests are most sensitive, i.e., among low emitters. Finally, let us mention that we have a strong agreement in the rankings between M1 and our flexible modeling under a series expansion in the IMR. The rank correlation with M1 is 0.96, all Jaccard indices exceed 0.60 and are as high as 0.88. Besides, the higher-order terms do not seem to matter much for the carbon premium computed next.

Given the sizeable cross-method discrepancies in correlations and Jaccard agreement, our next step is to assess the implications for estimating the carbon premium. Following Zhang (2025), we form - at the monthly frequency and for each emission estimator - two portfolios: High Emitters (top 30% by emissions) and Low Emitters (bottom 30%). We then report in Table 5 the average returns of the High and Low portfolios and the return differential (High - Low), which we interpret as the carbon premium. Comparing this spread across the five estimation methods reveals how

measurement choices map into inferences about the magnitude and potentially the sign of the carbon premium. Table 5 yields three main takeaways. First, using the vendor series, we recover a carbon premium of -0.36% (monthly), statistically significant at the 10% level, which mirrors prior evidence (e.g., Zhang (2025)) even though we adopt fiscal rather than effective timing to extend the sample, and we have two additional years as we end in 2023 instead of 2021, hence a helpful benchmark. Green stocks outperformed brown stocks like in Zhang (2025) reflecting the increased demand for greenness. Second, the sign of the premium is robust across all five estimators under value weighting: it is negative (i.e., low emitters outperform high emitters). The vendor series delivers magnitudes and significance similar to the naive size- and revenue-scaled imputations, with the latter often exhibiting slightly larger t-statistics. Third, relative to these three alternatives, our high-dimensional imputations, i.e. M1 and M2, imply a substantially larger premium in absolute value (same sign), underscoring that measurement and classification materially affect the economics. At the factor leg level, both high and low portfolios earn average excess returns above the market 1.05% over the same period; medians exceed means for all estimators, consistent with left-skewed return distributions (more pronounced among low emitters). Skewness and kurtosis estimates do not indicate extreme tail risk, reinforcing that the main differences across methods operate through firm classification rather than unusual higher-moment behavior. These patterns align with the Jaccard results, where the largest classification disagreements occur in the bottom deciles (10% and 30%). As already mentioned, the carbon premium remains stable under the flexible modeling at -0.47% (t(mean) = -2.39) instead of -0.49% (t(mean) = -2.55) under M1.

	High Emitters					**Low Emitters**					**High minus Low Emitters**				
	M1	**M2**	**Vendor**	**Size**	**Revenue**	**M1**	**M2**	**Vendor**	**Size**	**Revenue**	**M1**	**M2**	**Vendor**	**Size**	**Revenue**
Mean	1.31	1.31	1.41	1.38	1.40	1.80	1.83	1.77	1.75	1.73	-0.49	-0.52	-0.36	-0.37	-0.33
t(Mean)	4.26	4.24	4.49	4.45	4.42	4.54	4.56	4.46	4.59	4.63	-2.55	-2.63	-1.73	-2.33	-2.04
Median	1.55	1.54	1.74	1.73	1.63	2.27	2.25	1.85	2.24	2.06	-0.63	-0.47	-0.40	-0.19	-0.43
Std	3.98	4.00	4.07	4.01	4.12	5.14	5.21	5.16	4.93	4.85	2.49	2.58	2.73	2.06	2.09
Skewness	-0.09	-0.12	-0.13	-0.09	0.01	-0.15	-0.15	-0.11	-0.20	-0.16	-0.01	0.13	0.22	-0.18	-0.18
Kurtosis	3.50	3.51	3.54	3.70	3.72	3.38	3.45	3.57	3.13	3.48	2.76	3.00	2.95	2.85	2.83
VaR(5%)	-6.02	-6.03	-6.46	-6.09	-5.99	-8.05	-7.93	-7.65	-7.34	-7.30	-4.94	-5.21	-4.97	-4.41	-3.96
ES(5%)	-7.66	-7.78	-7.93	-7.73	-7.71	-9.83	-10.02	-9.81	-9.11	-9.03	-5.64	-5.80	-5.74	-4.85	-4.92

Table 5: Carbon Premium

To assess whether emissions-based sorting differences translate into differences in the types of firms selected into each bucket, Table 6 reports, for each estimation method, the monthly average of value-weighted carbon emissions and characteristics of the firms in each bucket. The characteristics include, size, book-to-market, profitability, investment, leverage, and revenue. Table 6 confirms the hierarchy documented above. The vendor series systematically underestimates emissions relative to our characteristics-based imputations: it is below M2 and even further below M1. The understate-

ment is especially pronounced among low emitters: the vendor average is about 4.89 thousands of tCO2e versus 11.76 under M1, whereas the gap is much smaller (in relative terms) among high emitters (about 10.84 versus 12.08 millions of tCO2e). By contrast, the naïve imputations based on size and revenue overestimate emissions by low emitters when benchmarked to M1: size implies 47.23 thousands versus 11.76 under M1 (9.56 under M2), and revenue implies about 22.56 versus 11.76. These patterns align with the Jaccard results, where the largest classification disagreements occur in the bottom deciles (10% and 30%).

	High Emitters					**Low Emitters**				
	M1	**M2**	**Vendor**	**Size**	**Revenue**	**M1**	**M2**	**Vendor**	**Size**	**Revenue**
	Carbon Emissions (in Thousands)									
Mean	12081.34	11671.26	10835.24	11872.30	13119.61	11.76	9.56	4.89	47.23	22.56
t(Mean)	29.43	29.83	31.63	29.52	29.93	20.00	20.31	16.32	25.70	39.24
Median	12387.13	11832.65	11725.31	10597.20	12647.28	10.16	8.53	6.27	46.32	23.58
Std	5320.33	5071.82	4439.48	5213.19	5682.46	7.62	6.10	3.89	23.82	7.45
Skewness	0.04	0.02	-0.01	0.60	0.14	0.46	0.41	-0.06	1.06	-0.22
Kurtosis	1.65	1.61	1.41	2.33	1.58	2.27	2.16	1.10	4.97	2.70
	Characteristics									
Size	18.05	18.06	18.09	18.00	18.02	17.13	17.01	16.21	17.83	17.44
Book-to-market	1.61	1.60	1.56	1.58	1.64	1.67	1.64	1.82	1.77	1.80
Profitability	0.18	0.18	0.19	0.18	0.17	0.20	0.19	0.16	0.20	0.19
Investment	0.07	0.07	0.08	0.09	0.08	0.17	0.18	0.18	0.14	0.14
Leverage	3.09	3.14	3.18	3.18	3.05	4.13	4.06	4.89	4.37	4.47
Revenue	55.91	55.76	55.98	52.93	54.53	17.66	13.86	6.88	38.53	26.73

Table 6: Carbon Factors Emissions and Characteristics

Firm characteristics comove with emissions in expected ways. High emitters are larger and generate higher revenues than low emitters, with revenue differences notably more pronounced than size differences. Book-to-market and profitability are broadly similar across buckets. Investment, measured by asset growth, is higher among low emitters, indicating more aggressive investment policies (relative to conservative firms in Fama and French (2015) sense). Leverage is modestly higher among low emitters. Overall, the most consequential cross-method differences concentrate among low emitters, consistent with where classification gaps are largest and where carbon premium tests are most sensitive.

5 Concluding thoughts

We address the issue of sample selection bias in the context of firm-level carbon emissions estimation, which has been largely overlooked in the literature. The economic cost of this bias is substantial, as it leads to significant underestimation of firm-level carbon emissions. We assert green silence to describe adverse selection in non-reporting firms that possess private information on their carbon emissions and use it to their benefit. Our approach not only quantifies the statistical and economic significance of this bias but also enables an empirical inference about the extent of green silence. To get consistency of variable selection for carbon estimation, we extend the two-step procedure of Heckman (1979) to a three-step procedure. In the theoretical side, we establish asymptotic normality of the estimated carbon regression parameter in the presence of sample selection. Such an asymptotic analysis decouples from Heckman (1979) because joint asymptotic analysis on the parameter of sample selection bias and nuisance parameters is impossible in the presence of the curse of dimensionality. Here, nuisance parameters are potentially biased from regularisation, and we need to rely on an extension exploiting a DML approach (Chernozhukov et al. (2018)).

Our empirical analysis reveals that sample selection substantially biases firm-level carbon emissions estimates, leading to understatements that distort both carbon tax revenue projections and social cost of carbon calculations. We anticipate that similar selection biases afflict other climate- and environment-related disclosures, from pollution-control investments to ecosystem impact assessments. By applying the correction methodology developed here, researchers and policymakers can mitigate these biases - thereby obtaining more accurate measures of firm pollution-control adoption, associated costs, and the true valuation of environmental externalities.

Studies show that carbon markets, most notably the EU Emissions Trading System (EU ETS), have provided measurable incentives for firms to reduce emissions (e.g., Colmer et al. (2025)). In the same spirit, expanding emissions inventories based on credible and transparent estimation can function as an informational disciplining device. Once firms realize their emissions will be imputed even without voluntary disclosure, opacity ceases to be a profitable strategy. Provided the estimates are not downward-biased, such inventories increase firm exposure to investors, lenders, customers, and watchdogs, activating reputational and financial channels and strengthening ex-ante accountability. In this way, inventories complement the carbon price signal by adding an information-based layer of incentives that extends pressure beyond formally covered entities and accelerates the adjustment of corporate behavior.

Carbon emissions are also pivotal for quantifying the effects of impact investing via the cost-of-capital channel. Recent studies - despite differing on the methodology for measuring the carbon

premium as well - rely almost exclusively on vendor-provided emissions data (e.g., Aswani et al. (2024), Bolton and Kacperczyk (2021, 2023), Zhang (2025)). Because these third-party estimates understate actual emissions (as documented in our empirical study), many firms are misclassified - appearing in the wrong "high-emitter" or "low-emitter" buckets. Consequently, long–short carbon factors designed to capture the carbon premium will mechanically mismeasure it. An accurate measure of the carbon premium is a prerequisite for disentangling investor varied motivations for engaging in impact investing (Starks (2023)).

References

Acemoglu, D., Akcigit, U., Hanley, D. and Kerr, W. (2016). Transition to clean technology, *Journal of Political Economy* **124**(1): 52–104.

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

Avramov, D., Lioui, A., Liu, Y. and Tarelli, A. (2025). Dynamic ESG equilibrium, *Management Science* **71**(4): 2867–2889.

Barnett, M., Brock, W. and Hansen, L. P. (2020). Pricing uncertainty induced by climate change, *The Review of Financial Studies* **33**(3): 1024–1066.

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*, Springer Science & Business Media.

Bia, M., Huber, M. and Laffér̀s, L. (2024). Double machine learning for sample selection models, *Journal of Business & Economic Statistics* **42**(3): 958–969.

Bolton, P. and Kacperczyk, M. (2021). Do investors care about carbon risk?, *Journal of Financial Economics* **142**(2): 517–549.

Bolton, P. and Kacperczyk, M. (2023). Global pricing of carbon-transition risk, *The Journal of Finance* **78**(6): 3677–3754.

Bond, P. and Zeng, Y. (2022). Silence is safest: Information disclosure when the audience's preferences are uncertain, *Journal of Financial Economics* **145**(1): 178–193.

Busch, T., Johnson, M. and Pioch, T. (2022). Corporate carbon performance data: Quo vadis?, *Journal of Industrial Ecology* **26**(1): 350–363.

Cai, Y. and Lontzek, T. S. (2019). The social cost of carbon with economic and climate risks, *Journal of Political Economy* **127**(6): 2684–2734.

Ceccarelli, M., Ramelli, S. and Wagner, A. F. (2024). Low carbon mutual funds, *Review of Finance* **28**(1): 45–74.

Chatterjee, A. and Lahiri, S. N. (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap, *The Annals of Statistics* **41**(3): 1232–1259.

Chen, A. Y. and McCoy, J. (2024). Missing values handling for machine learning portfolios, *Journal of Financial Economics* **155**: 103815.

Chen, A. Y. and Zimmermann, T. (2022). Open source cross-sectional asset pricing, *Critical Finance Review* **27**(2): 207–264.

Chen, X., Hansen, L. P. and Hansen, P. G. (2024). Robust inference for moment condition models without rational expectations, *Journal of Econometrics* **243**(1-2): 105653.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters, *The Econometrics Journal* **21**(1): C1–C68.

Cicala, S., Hémous, D. and Olsen, M. (2022). Adverse selection as a policy instrument: unraveling climate change, *NBER working paper* .

Colmer, J., Martin, R., Muûls, M. and Wagner, U. J. (2025). Does pricing carbon mitigate climate change? Firm-level evidence from the european union emissions trading system, *The Review of Economic Studies* **92**(3): 1625–1660.

Eggleston, H., Buendia, L., Miwa, K., Ngara, T. and Tanabe, K. (2006). 2006 IPCC guidelines for national greenhouse gas inventories.

EPA, U. (2023). EPA report on the social cost of greenhouse gases: Estimates incorporating recent scientific advances.

Exterkate, P. (2013). Model selection in kernel ridge regression, *Computational Statistics & Data Analysis* **68**: 1–16.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* **33**(1): 3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model, *Journal of Financial Economics* **116**(1): 1–22.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**(456): 1348–1360.

Flammer, C., Toffel, M. W. and Viswanathan, K. (2021). Shareholder activism and firms' voluntary disclosure of climate change risks, *Strategic Management Journal* **42**(10): 1850–1879.

Folini, D., Friedl, A., Kübler, F. and Scheidegger, S. (2025). The climate in climate economics, *The Review of Economic Studies* **92**(1): 299–338.

Fowlie, M., Reguant, M. and Ryan, S. P. (2016). Market-based emissions regulation and industry dynamics, *Journal of Political Economy* **124**(1): 249–302.

Freyberger, J., Neuhierl, A. and Weber, M. (2020). Dissecting characteristics nonparametrically, *The Review of Financial Studies* **33**(5): 2326–2377.

Gehricke, S., Leippold, M. and Schimanski, T. (2025). To disclose, or not to disclose: Evaluating the effectiveness of mandatory climate-related disclosure, *Available at SSRN 5246456* .

Gibbons, B. (2024). The financially material effects of mandatory nonfinancial disclosure, *Journal of Accounting Research* **62**(5): 1711–1754.

Hansen, L. P. and Scheinkman, J. A. (2025). Emission prices, biomass, and biodiversity in tropical forests, *AEA Papers and Proceedings* **115**: 391–396.

Hartzmark, S. M. and Sussman, A. B. (2019). Do investors value sustainability? A natural experiment examining ranking and fund flows, *The Journal of Finance* **74**(6): 2789–2837.

Heckman, J. J. (1979). Sample selection bias as a specification error, *Econometrica* **47**(1): 153–161.

Hong, Y. and White, H. (1995). Consistent specification testing via nonparametric series regression, *Econometrica* **63**(5): 1133–1159.

Huang, J., Horowitz, J. L. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models, *Annals of Statistics* **36**: 587–613.

Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models, *Statistica Sinica* pp. 1603–1618.

Huber, M. and Melly, B. (2015). A test of the conditional independence assumption in sample selection models, *Journal of Applied Econometrics* **30**: 1144–1168.

Ilhan, E., Krueger, P., Sautner, Z. and Starks, L. T. (2023). Climate risk disclosure and institutional investors, *The Review of Financial Studies* **36**(7): 2617–2650.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity, *Econometrica* **77**(5): 1481–1512.

Kremer, I., Schreiber, A. and Skrzypacz, A. (2024). Disclosing a random walk, *The Journal of Finance* **79**(2): 1123–1146.

Krueger, P., Sautner, Z., Tang, D. Y. and Zhong, R. (2024). The effects of mandatory ESG disclosure around the world, *Journal of Accounting Research* **62**(5): 1795–1847.

Lyubich, E. (2025). The role of people versus places in individual carbon emissions, *American Economic Review* **115**(5): 1439–1484.

Manski, C. (1989). Anatomy of the selection problem, *The Journal of Human Resources* **24**(3): 343–360.

Nath, I. (2025). Climate change, the food problem, and the challenge of adaptation through sectoral reallocation, *Journal of Political Economy* **133**(6): 1705–1765.

Net Zero, T. (2024). Net zero stocktake 2024: Assessing the status and trends of net zero target setting across countries, subnational governments, and companies, *New Climate Institute, Oxford Net Zero, Energy and Climate Intelligence Unit, and Data-Driven EnviroLab* .

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators, *Econometrica* **62**(6): 1349–1382.

Newey, W. K. (2009). Two-step series estimation of sample selection models, *The Econometrics Journal* **12**(suppl_1): S217–S229.

Newey, W. K., Powell, J. L. and Vella, F. (1999). Nonparametric estimation of triangular simultaneous equations models, *Econometrica* **67**(3): 565–603.

Pástor, L., Stambaugh, R. F. and Taylor, L. A. (2022). Dissecting green returns, *Journal of Financial Economics* **146**(2): 403–424.

Pastor, L., Stambaugh, R. F. and Taylor, L. A. (2025). Carbon burden, *Available at SSRN 4998860* .

Robinson, P. M. (1988). Root-n-consistent semiparametric regression, *Econometrica* **56**(4): 931–954.

Sautner, Z., Van Lent, L., Vilkov, G. and Zhang, R. (2023). Firm-level climate change exposure, *The Journal of Finance* **78**(3): 1449–1498.

Smola, A. J. and Schölkopf, B. (1998). *Learning with kernels*, Vol. 4, Citeseer.

Starks, L. T. (2023). Presidential address: Sustainable finance and ESG issues – value versus values, *The Journal of Finance* **78**(4): 1837–1872.

Stern, N. (2008). The economics of climate change, *American Economic Review* **98**(2): 1–37.

Stroebel, J. and Wurgler, J. (2021). What do you think about climate finance?, *Journal of Financial Economics* **142**(2): 487–498.

Tol, R. S. (2023). Social cost of carbon estimates have increased over time, *Nature climate change* **13**(6): 532–536.

Van den Bremer, T. S. and Van der Ploeg, F. (2021). The risk-adjusted carbon price, *American Economic Review* **111**(9): 2782–2810.

Vella, F. (1998). Estimating models with sample selection bias: A survey, *Journal of Human Resources* pp. 127–169.

Wooldridge, J. M. (1992). A test for functional form against nonparametric alternatives, *Econometric Theory* **8**(4): 452–475.

Yang, L., Lv, S. and Wang, J. (2016). Model-free variable selection in reproducing kernel hilbert space, *Journal of Machine Learning Research* **17**(82): 1–24.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **68**(1): 49–67.

Zhang, S. (2025). Carbon returns across the globe, *The Journal of Finance* **80**(1): 615–645.

Appendix: Proof of theorems

Proof of Theorem 2. Under Conditions 1- 5, we establish the proof for selection consistency of characteristics. By the Karush-Kuhn-Tucker (KKT) conditions, $\hat{b} \equiv \hat{b}(\hat{\theta})$ is the unique solution of (4) if we satisfy $-\mathbf{k}(j)^\top\big(Y - \mathbf{k}\hat{b}\big) + \lambda w_j \frac{\mathsf{K}_j \hat{b}_j}{\|\hat{b}_j\|_{\mathsf{K}_j}} = \mathbf{0}, \forall \hat{b}_j \neq \mathbf{0}$, and $\left\|-\mathbf{k}(j)^\top\big(Y - \mathbf{k}\hat{b}\big)\right\|_{\mathsf{K}_j,*} \leq \lambda w_j \|\mathsf{K}_j\|, \forall \hat{b}_j = \mathbf{0}$. Let $\mathbf{s}_h = w_h \operatorname{sgn}(b_h)$ for $h \in \mathsf{A}_b$, and $\mathbf{s}_1 = \mathsf{vec}\{\mathbf{s}_h; h \in \mathsf{A}_b\}$. Let $\hat{b}_1$ be the estimated active subset of $\hat{b}$

$$\hat{b}_1 = \big(\mathbf{k}_1^\top \mathbf{k}_1\big)^{-1}\big(\mathbf{k}_1^\top Y - \lambda_n \mathbf{s}_1\big) = b_{10} + \frac{1}{n}\Sigma_{\mathbf{k}_1}^{-1}\big(\mathbf{k}_1^\top \boldsymbol{\epsilon} - \lambda_n \mathbf{s}_1\big). \tag{26}$$

If $\hat{b}_1 =_s b_{10}$, then KKT condition holds for $\hat{b} = \big(\hat{b}_1^\top, \mathbf{0}^\top\big)^\top$. Since $\mathbf{k}\hat{b} = \mathbf{k}_1\hat{b}_1$ and $\mathbf{k}(j)$ are linearly independent for $j \in \mathsf{A}_b$, we deduce

$$\hat{b} =_s b_0, \quad \text{if} \begin{cases} \hat{b}_1 =_s b_{10}, \\ \left\|\mathbf{k}(j)^\top\big(Y - \mathbf{k}_1\hat{b}_1\big)\right\|_{\mathsf{K}_j,*} < \lambda_n w_j \|\mathsf{K}_j\|, & \forall j \neq \mathsf{A}_b. \end{cases} \tag{27}$$

Let $H_n = \mathbf{1}_n - \mathbf{k}_1\Sigma_{\mathbf{k}_1}^{-1}\mathbf{k}_1^\top/n$ be the projection to the oracle $\mathbf{k}_1^\top$. By (26), it follows that $Y - \mathbf{k}_1\hat{b}_1 = \boldsymbol{\epsilon} - \mathbf{k}_1\big(\hat{b}_1 - b_{10}\big) = H_n\boldsymbol{\epsilon} + \mathbf{k}_1\Sigma_{\mathbf{k}_1}^{-1}\mathbf{s}_1\lambda_n/n$, and by (27), we get

$$\hat{b} =_s b_0, \quad \text{if} \begin{cases} \operatorname{sgn}(b_{0j})(b_{0j} - \hat{b}_j) < |b_{0j}|, & \forall j \in \mathsf{A}_b, \\ \left\|\mathbf{k}(j)^\top\big(H_n\boldsymbol{\epsilon} + \mathbf{k}_1\Sigma_{\mathbf{k}_1}^{-1}\mathbf{s}_1\lambda_n/n\big)\right\|_{\mathsf{K}_j,*} < \lambda_n w_j \|\mathsf{K}_j\|, & \forall j \notin \mathsf{A}_b. \end{cases}$$

For $j \notin \mathsf{A}_b$, it suffices to show $\lim_{n\to\infty} \mathsf{P}[j \in \mathsf{A}_b] \to 0$. For $j \notin \mathsf{A}_b$, $\lambda_n w_j \to \infty$ as $n \to \infty$. By the adaptive irrepresentable condition in Condition 4, we know that $n^{-1}\left\|\mathbf{k}(j)^\top\mathbf{k}_1\Sigma_{\mathbf{k}_1}^{-1}\mathbf{s}_1\right\| = n^{-1}\sum_{h\in\mathsf{A}_b}\left\|\mathbf{k}(j)^\top\mathbf{k}(h)\Sigma_h^{-1}\mathbf{s}_h\right\|$ is bounded below η and $\eta < 1$, and $\mathbf{k}(j)^\top\big(H_n\boldsymbol{\epsilon}\big)$ converges to normality asymptotically. With condition (5) and $\lambda_n w_j \to \infty$, it completes the proof since $\lim_{n\to\infty}\mathsf{P}[j \in \mathsf{A}_b] = \mathsf{P}\left[\left\|\mathbf{k}(j)^\top\big(H_n\boldsymbol{\epsilon} + \mathbf{k}_1\Sigma_{\mathbf{k}_1}^{-1}\mathbf{s}_1\lambda_n/n\big)\right\| = \lambda_n w_j \frac{\mathsf{K}_j \operatorname{sgn}(\hat{b}_j)}{\|\hat{b}_j\|_{\mathsf{K}_j}}\right] \to 0$, provided that with a high probability $\lim_{n\to\infty}\mathsf{P}[j \notin \mathsf{A}_b] \to 1$. □

Proof of Theorem 3. We present a limiting distribution with selection bias correction. Let $b = b_0 + \frac{\boldsymbol{u}}{\sqrt{n}}$ where $\boldsymbol{u} = \sqrt{n} \times (b - b_0)$ and $\boldsymbol{u}_j \in \mathbb{R}^{L\times 1}$. $L_n(\boldsymbol{u}) = \frac{1}{2}\left\|\frac{1}{\sqrt{n}}\mathbf{k}\boldsymbol{\mu} + \epsilon\right\|^2 + \lambda_n n\sum_{j=1}^J w_j\big(b_{j0} + \frac{\boldsymbol{u}_j}{\sqrt{n}}\big)^\top\mathsf{K}_j\big(b_{j0} + \frac{\boldsymbol{u}_j}{\sqrt{n}}\big)$. Let $\hat{\boldsymbol{u}}_n = \arg\min L_n(\boldsymbol{u})$ such that $\hat{b} = b_0 + \frac{\hat{\boldsymbol{u}}_n}{\sqrt{n}}$, or $\hat{\boldsymbol{u}}_n = \sqrt{n}\big(\hat{b} - b_0\big)$. Define $V^{(n)}(\boldsymbol{u}) = L_n(\boldsymbol{u}) - L_n(\mathbf{0})$ and a decomposition of it, $V^{(n)}(\boldsymbol{u}) = V_1^{(n)}(\boldsymbol{u}) + V_2^{(n)}(\boldsymbol{u})$, where $V_1^{(n)}(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^\top\Sigma_{\mathbf{k}}\boldsymbol{u} - \frac{1}{\sqrt{n}}\boldsymbol{u}^\top\mathbf{k}^\top\epsilon$ with $\epsilon = \hat{\Gamma} + \varepsilon$ and $V_2^{(n)}(\boldsymbol{u}) = \sqrt{n}\lambda_n\sum_{j=1}^J\sqrt{n}w_j\mathsf{K}_j\left[\big(b_{j0} + \frac{\boldsymbol{u}_j}{\sqrt{n}}\big)^\top\big(b_{j0} + \frac{\boldsymbol{u}_j}{\sqrt{n}}\big) - b_{j0}^\top b_{j0}\right]$. Under Condition

5 and milder regularisation $\lambda_n/\sqrt{n} \to 0$, $\frac{\lambda_n}{\sqrt{n}} \sum w_j \mathsf{K}_j \left[\left(b_{j0} + \frac{\boldsymbol{u}_j}{\sqrt{n}}\right)^\top \left(b_{j0} + \frac{\boldsymbol{u}_j}{\sqrt{n}}\right) - b_{j0}^\top b_{j0} \right] = o_p(1)$ by Slutsky theorem for $b_{j0} \neq 0$, we get $V_2^{(n)}(\boldsymbol{u}) \xrightarrow{p} 0$. Define $\widehat{\Gamma} = \Gamma + \Delta_\Gamma$, $\Delta_\Gamma := \widehat{\Gamma} - \Gamma$, and decompose

$$\frac{1}{\sqrt{n}} \mathbf{k}^\top \widehat{\Gamma} = \underbrace{\frac{1}{\sqrt{n}} \mathbf{k}^\top \Gamma}_{M^{(0)}} + \underbrace{\frac{1}{\sqrt{n}} \mathbf{k}^\top \Delta_\Gamma}_{M^{(e)}}.$$

Hence $V_1^{(n)}(u) = \frac{1}{2} u^\top \Sigma_{\mathbf{k}} u - \frac{1}{\sqrt{n}} u^\top \mathbf{k}^\top \varepsilon - \frac{1}{\sqrt{n}} u^\top \mathbf{k}^\top \widehat{\Gamma} = \frac{1}{2} u^\top \Sigma_{\mathbf{k}} u - u^\top W - u^\top M^{(0)} - u^\top M^{(e)}$, with $W := \frac{1}{\sqrt{n}} \mathbf{k}^\top \varepsilon$. We impose the mild centring assumption $\mathbb{E}[\mathbf{k}_i \Gamma_i] = 0$. Indeed, by the solution of a unique minimiser of $L_n(\boldsymbol{u})$ and by the Lindeberg central limit theorem, we get $M^{(0)} = \frac{1}{\sqrt{n}} \sum_{i=1}^n k_i \Gamma_i \xrightarrow{d} \mathcal{N}\big(0,\ (\Gamma'_\Theta \Sigma_\Theta \Gamma'^\top_\Theta)\, \Sigma_{k_1}\big)$, where the variance $(\Gamma'_\Theta \Sigma_\Theta \Gamma'^\top_\Theta)$ arises from Proposition 2. By cross-fitting, the estimator $\widehat{\Gamma}$ is computed on auxiliary folds and then evaluated on the held-out observations. Hence, $\Delta_{\Gamma,i}$ is (approximately) independent of $\mathbf{k}_i$. If the first-stage error satisfies the rate $\max_{1 \le i \le n} |\Delta_{\Gamma,i}| = o_P(1)$, and $\frac{1}{n} \sum_{i=1}^n \|\mathbf{k}_i\|^2 = O_P(1)$ in Condition 7, then by Cauchy–Schwarz, we have $|M^{(e)}| = \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{k}_i \Delta_{\Gamma,i} \right| \le \sqrt{\frac{1}{n} \sum_{i=1}^n \|\mathbf{k}_i\|^2} \cdot \max_i |\Delta_{\Gamma,i}| = o_P(1)$. Such a rate is standard in the DML/cross-fitting literature: it suffices that $\widehat{\Gamma}$ attains a uniform consistency rate (or a suitable ℓ_2-rate) on auxiliary samples. Under $\mathbb{E}[\mathbf{k}_i \Gamma_i] = 0$ and the cross-fitting error control above, the two stochastic terms W and $M^{(0)}$ admit a joint Gaussian limit and $M^{(e)} = o_P(1)$. The subsequent epiconvergence argument then yields $\sqrt{n}(\widehat{b} - b_0) \xrightarrow{d} N\big(0,\ (\sigma^2 + \Gamma'_\Theta \Sigma_\Theta \Gamma'^\top_\Theta)\, \Sigma_{\mathbf{k}_1}^{-1}\big)$, as claimed. □