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

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

September 26, 2025

## Abstract

Voluntary carbon disclosure collapses into a paradox of green silence: firms choose to disclose emissions based on strategic incentives (e.g., correcting vendor overestimates), while high emitters may exploit vendor estimation bias. Mirroring Heckman sample selection bias, this self-censorship skews disclosed emissions into non-random samples, distorting climate risk pricing and policy. We bridge economic problem and machine learning, proposing a Heckman-inspired three-step framework in high-dimensional settings to correct for strategic non-disclosure and ensure variable selection consistency in the presence of sample selection bias. By integrating kernel group lasso (KG-lasso) and double machine learning (DML) from neighbouring firms, i.e., using information from carbon next door, we unveil systematic underestimation: empirical analysis of 3444 unique US firms (2010-2023) rejects the null of no selection bias. Our findings indicate that voluntary disclosure induces adverse selection, where green silence rewards polluters and undermines decarbonization. Underestimation translates to a \$2.6 billion shortfall in tax revenues and up to \$525 billion hidden social cost of carbon. Our high-dimensional imputations also imply a substantially larger carbon premium.

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

---

[†]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 reach new highs and climate impacts intensify globally, according to the UNEP Emissions Gap Report 2024.[1] Monitoring carbon emissions by firms is key to achieving a Net Zero Target (Net Zero (2024)). Firms disclose emissions voluntarily, and data vendors collect numbers. Amongst data vendors, the Carbon Disclosure Project (CDP) data is most widely used by academics and practitioners, and serves as the basis for other data vendors. According to the CDP disclosure report, 23,188+ firms disclosed voluntarily climate-related information in 2023, 140% increase from disclosure in 2020.[2] Out of these firms, 8000 (35%) disclosed for the first time. Yet, just under 400 companies (2%) were A listed by CDP, which is recognized for the very high quality of the published information.[3] While CDP coverage is on a voluntary basis, other data vendors select firms for their databases and provide estimates for non-disclosing companies. These estimated emissions account for around 2/3 of the data.

In the disclosure process, we likely face a sample selection issue inducing biased estimators (Heckman (1979)). When firms with superior carbon information strategically withhold data - anticipating that third-party estimates underestimate their true environmental impact - voluntary disclosure regimes morph into arenas of `green silence`, mirroring the economic notion of adverse selection. This self-censorship creates bias: reported carbon footprints become non-random samples, systematically skewed toward firms with fewer incentives to hide. Not surprisingly, corporate disclosure of environmental, social, and governance information has become a focal point for academics, practitioners, and regulators seeking to understand how transparency shapes market outcomes (Ilhan et al. (2023)). Yet, in the absence of uniform mandates, firm voluntary reporting behavior remains uneven: while long-term institutional shareholder activism can effectively compel firms to reveal climate-risk exposures - yielding measurable valuation premiums (Flammer et al. (2021)) - managerial uncertainty about stakeholder preferences and risk aversion sometimes induces strategic silence rather than full disclosure (Bond and Zeng (2022)). Recent regulatory innovations demonstrate that compulsory climate reporting not only elevates both the quantity and quality of firm disclosures but also reorients capital flows toward lower-carbon investments (Gibbons (2024), Gehricke et al. (2025)). Dynamic disclosure models further reveal that managers optimally release unfavorable information only below certain thresholds - sacrificing short-term price levels to reduce long-run valuation uncertainty (Kremer et al. (2024)) - and that well-designed ESG mandates enhance stock liquidity, particularly for firms with weaker preexisting information environments (Krueger et al. (2024)). Together, these find-

---

[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.

ings underscore the multifaceted drivers of corporate transparency and its critical role in promoting market efficiency and sustainable finance.

Sample selection bias in carbon disclosure arises not only from firm strategic self-censorship but also from data vendor reliance on incomplete information. Vendors lack granular data on small firms (e.g., limited public sustainability reports, opaque supply chains). Just as Heckman observed that wage samples exclude workers whose reservation wages keep them out of the labor force, carbon disclosures exclude firms whose unobserved environmental performance incentivized green silence (adverse selection in non-reporting firms which possess private information on their carbon emissions and use it to their benefit). Reported emissions data - like the earnings of migrants or trainees - do not reflect the counterfactual: what nondisclosers would have reported if compelled to transparency. Conventional comparisons (disclosers vs. nondisclosers) thus misestimate the true 'treatment effect' of decarbonization policies, much as uncorrected wage studies misestimate the value of union membership. Only by modeling the selection process itself - why firms opt out of disclosure - can we disentangle green rhetoric from sustainability.

We establish the parallel to Heckman selection bias. In Heckman seminal work, sample selection bias arises when individuals self-select into a study (e.g., migrants, union workers). Similarly, in voluntary carbon disclosure: *(1) self-selection by firms*: firms choose to disclose emissions based on strategic incentives (e.g., correcting vendor overestimates or avoiding scrutiny); *(2) vendor estimation bias*: third-party vendors estimate emissions for non-disclosers using incomplete or skewed data, mirroring Heckman analyst-driven selection. By addressing self-selection bias, Heckman approach indirectly aids in analysing markets plagued by adverse selection. 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 (adverse selection in non-reporting firms) within the carbon disclosure landscape.

While the IPCC Guidelines set a clear method for differentiating between "sectors of economy" (Eggleston et al. (2006)), these sectors are quite different from those understood by economists. In the IPCC Guidelines, a sector is a grouping of activities, while in economics, a sector is a grouping of similar economic actors. The energy sector, for example, includes most combustion of energy, whether the activities are undertaken by enterprises whose main activity is energy production or not. All household combustion of gasoline in private transportation is included in the energy sector, whereas under economic accounts, such activities are 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 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.

The carbon estimates provided by data providers, including CDP, *Trucost*, MSCI, Sustainalytics, Thomson Reuters, Bloomberg, and ISS often diverge due to differences in the definitions of nearest peer groups. For instance, Thomson Reuters employs a carbon-to-size ratio as the matching criterion to identify potential peer companies in the same industry if there are at least 10 firms, and the firms are extended to industry group, business sector, and finally economic sector, following that order until there are sufficient observations. Unlike Thomson Reuters, MSCI uses total revenue rather than size as the primary matching criterion for identifying nearest peers. Beyond firm size and revenue, other firm characteristics such as total assets, total sales, number of employees, and net property and equipment values can serve as proxies for the scale of operations, which in turn predict carbon output. These variations in the chosen approach raise a fundamental question: Which firm characteristics are most crucial for identifying the nearest peer group to infer undisclosed carbon output? Can we trust those selected firm characteristics and the estimated carbon emissions in the presence of sample selection? Can we simply rely on naïve imputation based on size and revenue? Research publications using vendor estimates by MSCI ESG, Refinitiv, Sustainalytics, and Trucost have experienced a huge growth from a dozen per year in 2008 to several thousands in recent years, according to the Dimensions research database. A key question is to check potential biases in those estimates.

We propose kernel group lasso (KG-lasso) to identify carbon neighbours across characteristics-sorted portfolios that mimic activity-based sectors. The group lasso framework is employed to select groups of portfolios sorted by key characteristics that are most informative for deriving carbon insights. Hence, we aim to leverage information from carbon next door, i.e., neighbouring firms. The kernel function generates a vector of weights for the sorted portfolios, indicating the similarity between an undisclosing firm and the $L$ portfolios in terms of the selected characteristics. By leveraging a feature map, the kernel function measures similarity in the feature space rather than the original characteristics space. A significant advantage of this approach is that the kernel captures nonlinearity and high-order interactions, extending beyond simple linear correlations. Concerning the identification issue in the sample selection problem, we propose an adaptive KG-lasso to differentiate the variable selection contributors from the sample selection contributors. The estimates from the selection equation are adaptive weights used to differentiate regularisations at the group level, hence the resulting active set in the variable selection equation and the active set in the sample selection equation have a bounded intersection. Ultimately, the adaptive weights benefit exclusion restrictions.

Drawing on Heckman remedy for sample selection bias, we shed light on gaps in carbon reporting and show how accounting for strategic non-disclosure can better align incentives and strengthen

decarbonization policies. In the presence of high-dimensional firm characteristics that determine firm heterogeneity, we cannot use off-the-shelf penalisation techniques available in the literature. 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 equation and variable selection equation. It is stringent to consider advanced modern approaches to circumvent the curse of dimensionality in our framework. Hence, on the theoretical side, we contribute to the literature by showing i) the asymptotic distribution of a test statistic for sample selection in the presence of high-dimensional nuisance parameters, ii) asymptotic consistency of variable selection in the carbon function after sample selection bias correction, iii) the asymptotic normality of the estimated carbon regression parameter in the presence of sample selection. To get i), we rely on the recent double machine learning (DML) approach by Chernozhukov et al. (2018). The advantage of DML leverages the Neyman orthogonality to make the parameter of sample selection bias insensitive to inconsistency in the high-dimensional nuisance estimates. The inconsistency arises from the regularisation bias from both nuisance estimates. As long as the coefficient of selection bias can be consistently estimated, we can do "post" variable selection in the variable selection equation to attain the consistent estimators and variable selection. One of the by-products is to deliver doubly robust score test for sample selection bias. The existing tests may fail to have unit power asymptotically against a wide range of regularisation bias in variable selection equation and sample selection equation.

For consistency of variable selection, we extend the two-step procedure of Heckman (1979) to a three-step procedure.[4] The first step is to estimate nuisance parameters and plug-in these nuisance parameters into the main equation to consistently estimate parameter of sample selection bias in the second step using DML approach to get i). In the last step for post variable selection, we consistently estimate the main equation and derive consistent variable selection to get ii). This three-step procedure generalises Heckman (1979) to a high-dimensional setup. In addition to estimation strategies, we establish asymptotic analysis in our framework for iii). The asymptotic analysis in the proposed framework decouples from Heckman (1979) because joint asymptotic analysis on parameter of sample selection bias and nuisance parameters is impossible in the presence of the curse of dimensionality. Indeed, nuisance parameters are potentially biased from regularisation. We need to first asymptotically analyse the parameters of sample selection bias, and the parameters in the sample selection equation separately, then given the studied asymptotic properties, we can finally analyze asymptotically the estimated parameters in the variable selection equation for estimation consistency and variable selection consistency.

Our approach à la Heckman exploits a plug-in of a bias correction term. Hence, the proposed

---

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

framework takes a different route from the control function approach by 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 focus is twofold: a) establishing the consistency of the coefficient associated with the bias correction term, and b) deriving the asymptotic normality of the estimated parameters in the variable selection equation. This focus yields two new econometric outcomes: a doubly robust score test for detecting sample selection bias, and consistent variable selection in the presence of such a bias.

To quantify the impact of selection bias on carbon emissions estimates, we use annual carbon data from *Trucost* covering 3444 unique US firms and 22,043 firm-year observations over the period January 2010 to December 2023. A substantial share of these data points is estimated by the vendor rather than disclosed by firms. We leverage a rich set of firm characteristics (173 in total) both for sample and variable selection. Our primary finding 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 the unobserved determinants of selection and outcome equations. Unobserved factors that increase the likelihood of voluntary disclosure are negatively associated with unobserved drivers of carbon emissions estimates. Firms with greener unobserved attributes are more likely to disclose, seeking to avoid vendor overestimation, and also tend to generate lower emissions due to their climate-conscious behavior. Ignoring selection bias results in a substantial underestimation of scope 1, 2, and 3 emissions when considered separately. Encouragingly, we document a steady decline in this bias over the sample period, particularly for scope 1 and 2 emissions. This trend appears related to the expanded data coverage following *Trucost* acquisition by S&P in 2016.

A second key finding concerns the role of firm characteristics in the selection process. Sample selection is primarily driven by indicators of firm quality, with firm size, age, and trading volume emerging as dominant predictors. In contrast, the variables selected in the variable selection equation are more closely related to firm future growth opportunities - such as R&D intensity, profitability, and investment activity - as well as capital structure. Notably, firm size plays no significant role in the variable selection stage. Only a small number of characteristics are inactive across both steps of our methodology for scope 1. The number of characteristics for scope 2 and 3 are higher even if less than 10. Prominent characteristics also differ across different scope emissions. Common ones include debt issuance, R&D, and volatility, yet each scope has its own key drivers. These findings underscore the empirical relevance of a high-dimensional approach in both sample and variable selection. In contrast, data vendors often rely on a limited set of firm characteristics to impute emissions, which is likely to result in substantial underestimation - a pattern that is readily verified empirically. To highlight the importance of dimensionality, we also implement our method using only firm size or

revenue as predictors. In these restricted specifications, the magnitude of the sample selection bias is several orders of magnitude larger than when employing the full set of characteristics. This raises concerns for studies aiming to measure the carbon premium or carbon burden, as many rely on only a small number of characteristics (see, e.g., Aswani et al. (2024), Bolton and Kacperczyk (2021, 2023), Pastor et al. (2025), Zhang (2025)).

A third finding concerns the carbon tax revenue shortfall implied by underestimated emissions. Using our high-dimensional approach and accounting for underestimation across all emission scopes, we estimate a conservative tax revenue loss of $2.65 billion. This estimate is likely understated, as the methodology used by data vendors to impute emissions is typically undisclosed. Comparing vendor-imputed emissions to our high-dimensional predictions that correct for selection bias, we observe substantial underestimation on the part of the vendor. This discrepancy translates into a potential tax revenue shortfall exceeding $9 billion. To the credit of the data vendor, we note that the pattern of decreasing selection bias over time - previously documented using our own high-dimensional estimates - also holds in the vendor data. Moreover, underestimation based on simplified imputations using only firm size or revenue is considerably larger than that observed in the vendor estimates. This suggests that while vendor-based estimates fall short relative to a high-dimensional correction, they still outperform naïve low-dimensional approaches commonly used in practice. A forward-looking perspective on our findings can be gained by considering the social cost of carbon. The social cost of carbon represents the present value of the estimated economic damages caused by the emission of one additional ton of carbon dioxide. It serves as a key benchmark in evaluating the benefits of emissions reductions and is widely used in climate policy and cost-benefit analysis. The US Environmental Protection Agency (EPA) periodically publishes estimates of this cost (see EPA (2023)). Folini et al. (2025) conduct an extensive analysis of the sensitivity of the social cost of carbon to model specification and parameter calibration. A critical input in such assessments of the global social cost of carbon is the measurement of firm emissions. Based on our corrected emissions estimates, we find that the implied economic cost of underreported carbon emissions could be as high as $525 billion. This figure vastly exceeds the estimated tax revenue shortfall and underscores the broader societal implications of inaccurate carbon reporting.

Our fourth finding pertains to the impact of selection bias on carbon-premium measurement. Even if imperfect, there is solid evidence that vendor rankings are taken seriously by investors. Studies such as Hartzmark and Sussman (2019) and Ceccarelli et al. (2024) show how labels such as "sustainable" or "low carbon risk" drive fund flows. When vendor data understate the emissions of non-disclosers, firms are misclassified between "low" and "high" emitter portfolios, biasing both investor and mutual fund allocations and, ultimately, the estimated premium. We show that this misclassification is especially pronounced for vendor-, size-, and revenue-scaled imputations, com-

pared to our M1 (selection-corrected) and M2 estimators, which exploit a broader information set. The latter deliver a markedly larger carbon premium (in absolute value). These findings have direct implications for regulators when assessing the cost-of-capital channel of carbon risk.

The paper is organized as follows. In Section 2, we outline our model based on a high-dimensional regression with reproducing kernels. The model is made of a variable selection equation and a sample selection equation. We discuss the identification issues underlying our approach. In Section 3, we explain how to build a doubly robust score test for our sample selection model. We deploy the DML approach by Chernozhukov et al. (2018) to correct for regularisation bias. In Section 4, we explain how our kernel group lasso brings consistency of variable selection under sample selection bias. In Section 5, we describe the data and our empirical results. In the Appendix, we provide proofs of our theorems. In an Online Appendix, we gather additional Tables, and Figures and give 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

---

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

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 finance literature, portfolio-sorting is popular in return prediction as it flexibly exploits cross-sectional information. 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 $\ell$ 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 $k(.,.)$ 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 empirical exercise. The corresponding coefficient $b_{j,\ell}$ weights the kernel function. In (1), $b_{j,\ell}$ is the average carbon of $\ell$-th portfolio sorted by $j$. We can think of $k(X_{ij}, P_{j,\ell})$ as a smooth extension of the indicator function $\mathbf{1}(X_{ij} \in P_{j,\ell})$ to belong to $P_{j,\ell}$. 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 propose to replace the indicator $\mathbf{1}(X_{ij} \in P_{j,\ell})$ by the kernel function $k(X_{ij}, P_{j,\ell})$ to measure similarity between the two entries in a nonlinear fashion. 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 columns 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

---

[6]See Appendix OA.2 for more details.

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 linear regression based on reproducing kernels with many regressors $\mathbf{k}_i$ and many parameters in $b$. 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 $L$.

## 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 $Z_i$:

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

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

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

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

---

[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%.

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

$$\mathsf{E}[\epsilon_i|X_i, Z_i, D_i = 1] = \theta h(Z_i^\top \beta) = \theta h_i, \tag{7}$$

where $h(z) = \phi(z)/\Phi(-z)$ is known as the inverse Mills ratio, $h_i := h(-Z_i\beta)$. For consistency in carbon estimation, it is stringent to incorporate (7) into (4) for bias correction. We rewrite the primary equation as,

$$Y_i = \mathbf{k}_i b + \theta h_i + \epsilon_i, \tag{8}$$

where $\theta$ is proportional to the covariance between $\epsilon$ and $\mathsf{v}$, denoted by $\sigma_{\epsilon,\mathsf{v}}$.

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 $\theta$ insensitive to inconsistency in the high-dimensional nuisance estimates $b$ and $\beta$. The inconsistency arises from the regularisation bias in the nuisance estimates. As long as $\theta$, 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 $\beta$, and plug in these nuisances into (8) to consistently estimate $\theta$ 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 $\theta$, $\beta$ and $b$ is not possible in the presence of the curse of dimensionality. The estimators of $\beta$ 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 $\mathsf{E}_N$ denotes the sample mean of $N$ observations, $\Lambda_i(.)$ is the negative log-likelihood for the probit model evaluated at $i$, and $\left\|\cdot\right\|_1$ 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 $\lambda > 0$:

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

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 $\left\|b_j\right\|_{\mathsf{K}_j} = \left(b_j^\top \mathsf{K}_j b_j\right)^{1/2}$ , where $\mathsf{K}_j$ 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 $\mathsf{K}_j$. The norm $\left\|b_j\right\|_{\mathsf{K}_j}$ 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 $\ell$ and $\ell'$, sorted by $j$. If all entries in $\mathsf{K}_j$ 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 $\left\|b_j\right\|_{\mathsf{K}_j}$ 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 $\boldsymbol{z}_j = \mathbf{k}(j)^\top (Y - \mathbf{k}b_{-j})$, $b_{-j} = (b_1, \cdots, b_{j-1}, \mathbf{0}, b_{j+1}, \cdots, b_J)^\top$ and

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

Let $\mathbf{I}_L$ be $L$-dimensional matrix of ones. A closed-form solution for (11) is

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

## 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 inverse Mills ratio) which appears in the variable selection equation. However, as addressed by Vella (1998), the inverse Mills ratios 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.** $\epsilon_i$ and $\mathsf{v}_i$ are i.i.d. jointly normally distributed with zero means and finite second moment $\sigma^2$ 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 $\mathsf{A}_b$ and $\mathsf{A}_\beta$. As sparsity condition, we assume $|\mathsf{A}_b| < J$ and $|\mathsf{A}_\beta| < q$ and $q < p$.

**Assumption 4.** *Denote $\mathsf{P}^d$ the power set of $d$ and $d = \min(|\mathsf{A}_b|, |\mathsf{A}_\beta|)$. For an intersecting family $\mathcal{A} \subseteq \mathsf{P}^d$, and $1 \le s \le n$ 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 inverse Mills ratio 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 inverse Mills ratio, 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 $\gamma \ge 1$. 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 $w_j = \left(1 + |\hat{\beta}_j|\right)^\gamma$, $\gamma \ge 1$ 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 the excess of one. Clearly, the adaptive-weight regularisation 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.

# 3 Doubly robust score test for sample selection model

## 3.1 Neyman orthogonal score and asymptotic normality

To rigorously pin down the asymptotic theorem for $\theta$ and to test sample selection bias, we incorporate Neyman orthogonality conditions and derive the Neyman orthogonal score to make $\theta$ insensitive

---

[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.

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 $W_i = (Y_i, D_i, Z_i, X_i)$. 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 $\ell(W_i; \theta, \eta) = -\frac{1}{2}(Y - \mathbf{k}_i b - \theta h_i)^2$. We present the Neyman orthogonal score $\psi(W; \theta, \eta) = \partial_\theta \ell(W; \theta, b) - \mu \partial_b \ell(W; \theta, b)$, where $\eta = (b, h)$ and $h := h(Z\beta)$. $\mu$ solves the equation $\mathcal{J}_{\theta\eta} - \mu \mathcal{J}_{\eta\eta} = 0$, where $\mathcal{J}$ stands for Jacobian 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 $\partial_\eta \mathsf{E}[\psi(W; \theta_0, \eta_0)][\eta - \eta_0] = 0$. 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

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

Equation (15) forms a linear score (linear in $\theta$) to benefit from computational advantages to side step Jacobin matrix computation. We can decompose the entire score into two parts,

$$\psi(W_i; \theta, \eta) = \psi_a(W_i; \eta)\theta + \psi_b(W_i; \eta), \tag{16}$$

where $\psi_a(W_i; \eta) = -h_i(h_i - \mathsf{E}[h])$ and $\psi_b(W_i; \eta) = (Y_i - \mathbf{k}_i b)(h_i - \mathsf{E}[h])$. For $\theta = \theta_0$, we have a moment condition satisfying

$$\mathsf{E}[\psi(W_i; \theta_0, \eta_0)] = 0. \tag{17}$$

To satisfy the orthogonality conditions, following Chernozhukov et al. (2018) we deploy a $K$-fold cross-fitting algorithm 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 $\eta$ from auxiliary samples will be plugged into the score function to solve the moment condition for $\theta$ estimate. Given the efficient orthogonal score, $\hat{\theta}$, the average of $\theta$ estimates from $K$ folds, is the consistent estimator of $\theta_0$, as stated in the next theorem following directly from Chernozhukov et al. (2018) and based on the implementation algorithm:

15

---

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

---

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

**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 $\beta$ in one subsample, and $b$ in the other subsample to predict $\beta_k$ and $b_k$ in $\mathcal{W}_k$.

**3** Estimate $\theta_k$ using the plug-in $\eta_k = \{\beta_k, b_k\}$ under the moment condition in (17).

**4** Average $\theta_k$ across all $K$ subsamples to obtain $\hat{\theta} = \frac{1}{K} \sum_1^K \theta_k$.

**Output:** $\hat{\theta}$

---

**Theorem 1.** *If the score is the Neyman orthogonal efficient, and under the condition that the variance of the score $\psi$ is no-degenerate: all eigenvalues of matrix $\mathsf{E}[\psi(W; \theta_0, \eta_0)\psi(W; \theta_0, \eta_0)^\top]$ are bounded from below by $\tau_N > 0$ and suppose $\delta_N \geq N^{-1/2}$, then we have*

$$\sqrt{N}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \bar{\psi}(W_i) + O_p(\tau_N) \rightarrow \mathcal{N}(0, \sigma_\psi^2), \tag{18}$$

*where the reminder term requires to $\tau_N \leq \delta_N$, and $\bar{\psi}(W_i) := -\mathcal{J}_0^{-1}\psi(W_i, \theta_0, \eta_0)$ is the influence function, and Jacobian matrix $\mathcal{J}_0 = \partial_\theta \mathsf{E}[\psi(W, \theta, \eta_0)]\big|_{\theta=\theta_0}$, while the asymptotic variance is $\sigma_\psi^2 := \mathcal{J}_0^{-1}\mathsf{E}_N[\psi(W, \theta_0, \eta_0)\psi(W, \theta_0, \eta_0)^\top](\mathcal{J}_0^{-1})^\top$.*

We can replace $\sigma_\psi^2$ 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 (18) 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. Concerning these risks, we can impose the necessary regularity conditions to ensure the validity of Theorem 1. These regularity conditions include the near-orthogonality condition, a linear score function, a smooth score, the minimal singular value of Jacobian matrix and the statistical rate of risk associated with the score approximation in the presence of estimation risk of nuisance parameters. These regularity conditions are standard in the literature of M-estimators in the presence of high-dimensional nuisances. 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.

## 3.2 Doubly robust Score test for sample selection bias

We take a significant step further to establish the score test statistics based on the Neyman orthogonal score for a doubly robust sample selection bias test. The score test is based on the efficient score

criterion, and it has the advantage compared with other large sample tests, such as the likelihood ratio and Wald tests, of requiring estimation only under the null hypothesis. The score test, or LM test, is not new for sample selection bias. The earlier work by Melino (1982) has formalised a correspondence between the LM test and the t-test proposed by Heckman (1979). 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.

To isolate the impact from bias, we propose a doubly robust score test that replaces the efficient score vector with the Neyman orthogonal score. One merit of the doubly robust score test is its power in the presence of bias, hence, one can pay less effort on striking a balance between the bias term and the variance term, which is the key in the specification test literature. Wooldridge (1992) uses the bias to determine the limit distribution by controlling the variance so as to be negligible. In contrast, Hong and White (1995) use the variance to determine the limit distribution by controlling the bias so as to be negligible. As orthogonalization eliminates the bias term, we, therefore, focus on maximising the information of the variance of the score that exclusively determines the power of the test statistics. The doubly robust score test for the null $\theta = 0$ relies on the test statistic given $n$ samples denoted as $W_n$:

$$s_n := \psi(W_n); \theta, b)^\top \mathcal{V} \psi(W_n; \theta, b) \xrightarrow{d} \chi_1^2, \tag{19}$$

where $\mathcal{V} := \mathcal{H}_{\theta\theta} - \mathcal{H}_{\theta b} \mathcal{H}_{b\theta} (\mathcal{H}_{bb})^{-1}$ and $\mathcal{H} = -\partial^2_{(\theta' b')} \mathsf{E} \big[ \partial^2_{(\theta' b')'} \ell (W; \theta, b) \big]$ is the Hessian matrix $\begin{bmatrix} \mathcal{H}_{\theta\theta} & \mathcal{H}_{\theta b} \\ \mathcal{H}_{b\theta} & \mathcal{H}_{bb} \end{bmatrix}$.

We evaluate the score $\psi(W_n; \theta, b)$ 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 $\theta$. From standard results on score tests, the test statistic is asymptotically distributed as a chi-square distribution with one degree of freedom. Using Taylor expansion of the Lagrangian and information matrix equivalence, Aitchison and Silvey (1958) formally prove a chi-square limiting distribution. Compared to the Wald-type or $t$ test in Heckman (1979), the score test has a computational advantage because there is no need to solve for the moment condition in (17). In the case of a large sample, the power of test statistic $s_n$ is asymptotically equivalent to the likelihood ratio test (Silvey (1959)) for which we have consistency of the testing procedure.

17

# 4 Variable selection under sample selection bias

## 4.1 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 $\theta$ 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 $\theta$ in Theorem 1, we obtain a post-selection estimator in the presence of sample selection bias

$$\hat{b}_j(\hat{\theta}) = \left( \mathbf{I}_L - \lambda w_j \frac{\mathsf{K}_j}{\sqrt{\boldsymbol{z}_j(\hat{\theta})^\top \mathsf{K}_j \boldsymbol{z}_j(\hat{\theta})}} \right)_+ \boldsymbol{z}_j(\hat{\theta}). \tag{20}$$

Equation (20) 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 $\lambda$. 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.

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

$$-\mathbf{k}(j)^\top \big( Y - \mathbf{k}\hat{b} \big) + \lambda w_j \frac{\mathsf{K}_j \hat{b}_j}{\left\| \hat{b}_j \right\|_{\mathsf{K}_j}} = \mathbf{0}, \qquad \forall \ \hat{b}_j \neq \mathbf{0}, \tag{21}$$

$$\left\| \mathbf{k}(j)^\top \big( Y - \mathbf{k}\hat{b} \big) \right\| < \lambda w_j \left\| \mathsf{K}_j \right\|, \qquad \forall \ \hat{b}_j = \mathbf{0}. \tag{22}$$

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 car-

bon 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 $b_0 = (b_{10}, \cdots, b_{J0})$. Consistency in characteristics selection ensures a separation between the oracle active set $\mathsf{A}_b = \{j : b_{j0} \neq 0\}$ and the complement $\mathsf{A}_b^c = \{j : b_{j0} = 0\}$. We write $b_0 = (b'_{10}, b'_{20})'$, two subsets where $b_{10}$ is the subset with the indices of elements in $\mathsf{A}_b$ and $b_{20}$ is the subset with the indices of elements in $\mathsf{A}_b^c$. The size of $\mathsf{A}_b$ is quantified by its cardinality $|\mathsf{A}_b| = J_1$ and by $|\mathsf{A}_b^c| = J_2$ for $\mathsf{A}_b^c$, 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 $j \notin \mathsf{A}_b$, $c_n = \frac{\max_{j \notin \mathsf{A}_b} \|\mathbf{k}(j)^\top Y\|}{\left(\sum_{h \in \mathsf{A}_b} \|\mathbf{k}(h)^\top Y\|\right)/J_1} < 1$. Partial orthogonality implies $n^{-1}\lambda_{max}(\Sigma_{jh}) \leq \rho_n$, $j \notin \mathsf{A}_b, h \in \mathsf{A}_b$. 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 $j \notin \mathsf{A}_b$ the irrepresentable condition requires $\frac{1}{n\sqrt{L}}\sum_{h \in \mathsf{A}_b}\|\mathbf{k}(j)^\top \mathbf{k}(h)\Sigma_h^{-1}\mathbf{s}_h\| \leq \frac{c_n\rho_n}{d_n} \leq \eta$.*

**Condition 5** (**Bounded eigenvalue of kernel matrix**). *$\mathsf{K}_j$ is a symmetric $L \times L$ positive definite kernel matrix, with entries $[\mathsf{K}_j]_{\ell,\ell'} = \frac{1}{L}k(P_{j,\ell}, P_{j,\ell'})$ 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 $\rho_n$ 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, $\rho_n$ 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 $\mathrm{sgn}(\hat{b}) = \mathrm{sgn}(b_0)$.

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

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.

## 4.2    Asymptotic analysis under sample selection bias

The presence of sample selection bias poses a challenge to asymptotic analysis on the variable selection equation as the resulting limiting distribution is inconsistent. 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 sample selection equation and asymptotic normality of estimated bias term $\hat{\theta}\hat{h}$ where $\hat{h} := h(Z^\top \hat{\beta})$. 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 $\theta$ 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 $\lambda$. Let $\beta_0 = (\beta_{10}^\top, \beta_{20}^\top)^\top$ be a true coefficient with the true nonzero coefficient $\beta_{10}$ and true zeros captured by $\beta_{20}$. Denote the score function $\phi(Z_i, \beta_{10}) \equiv \partial_\beta \tilde{\Lambda}(Z_i, \beta_{10})$ and Jacobian matrix $\mathcal{J}_{\beta_{10}} = \partial_{\beta_1} \mathsf{E}\big[\phi(Z, \beta_1)\big]\big|_{\beta_1 = \beta_{10}}$ to arrive at the influence function $\bar{\phi}(Z_i) := -\mathcal{J}_{\beta_{10}}^{-1}\phi(Z_i\beta_{10})$. 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 $\beta_0$. Following their regularity conditions, we obtain the asymptotic normality of $\beta_1$ if the irrelevant covariates are known:

$$\sqrt{N}\big(\hat{\beta}_1 - \beta_{10}\big) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \bar{\phi}(Z_i) \to \mathcal{N}(0, \sigma_\phi^2), \tag{23}$$

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, $\sigma_\phi^2$ 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)

$$\Gamma(W_i; \theta, \beta) = \mathsf{E}\big[\epsilon_i\big|X_i, Z_i, D_i = 1\big] = \theta h(Z_i^\top \beta) = \theta h_i. \tag{24}$$

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 $\Gamma(\Theta + \ell) - \Gamma(\Theta) = \Gamma'_\Theta(\ell) + \mathbf{o}(\|\ell\|)$, $\ell \to 0$. The derivative map $\ell \to \Gamma'_\Theta(\ell)$ is matrix multiplication by the matrix $\Gamma'_\Theta = -\partial_{(\theta', \beta')} \mathsf{E}[\partial_{(\theta', \beta')'} \Gamma(W; \theta, \beta)] = \begin{bmatrix} \Gamma'_{\theta\theta} & \Gamma'_{\theta\beta} \\ \Gamma'_{\beta\theta} & \Gamma'_{\beta\beta} \end{bmatrix}$. Denote the asymptotic covariance matrix for $\Theta = (\theta, \beta)$,

$$\Sigma_\Theta = \begin{bmatrix} \sigma_\psi^2 & \Omega \\ \Omega & \sigma_\phi^2 \end{bmatrix}, \tag{25}$$

where $\Omega = \mathcal{J}_0^{-1} \mathsf{E}_N[\psi(W, \theta_0, \eta_0) \otimes \phi(Z, \beta_{10})^\top](\mathcal{J}_{\beta_{10}}^{-1})^\top$ 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 $\beta$. If there exists a linear derivative map $\Gamma'_\Theta$, the estimate of $\Gamma$ has asymptotic normality*

$$\sqrt{N}(\hat{\Gamma} - \Gamma) \xrightarrow{d} \mathcal{N}(0, \Gamma'_\Theta \Sigma_\Theta (\Gamma'_\Theta)^\top), \tag{26}$$

*based on the asymptotic covariance defined in (25).*

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 $\mathbf{k}_1 \in \mathbb{R}^{n \times J_1 L}$ 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 $J_1 L \times J_1 L$ 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.

**Theorem 3 (Asymptotic normality of characteristics estimate).** *Under Conditions 5 and 6 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 (23), the adaptive KG-lasso estimates $\hat{b} = (\hat{b}_j)_{j=1,\cdots,J}$ with $\hat{b}_j := \hat{b}_j(\hat{\theta})$ in (20) are consistent and possess oracle properties*

$$\sqrt{n}(\hat{b} - b_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\sigma^2 + \Gamma'_\Theta \Sigma_\Theta (\Gamma'_\Theta)^\top) \Sigma_{\mathbf{k}_1}^{-1}\right). \tag{27}$$

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 (19) is rejected. We obtain an augmented variance induced by $\left(\Gamma'_\Theta \Sigma_\Theta (\Gamma'_\Theta)^\top\right) \Sigma_{\mathbf{k}_1}^{-1}$. The proof of Theorem 3 is provided in the appendix.

# 5 Empirical results

## 5.1 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

---

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

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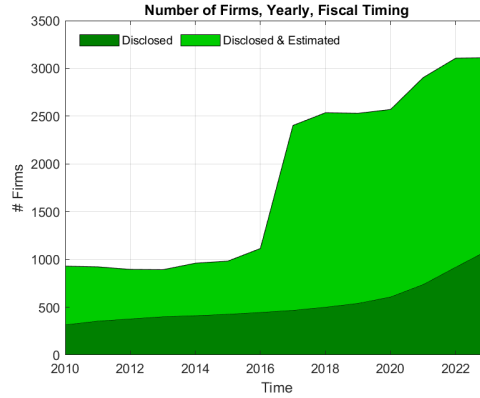


Figure 1: Number of firms over the sample period 2010 to 2023.

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.
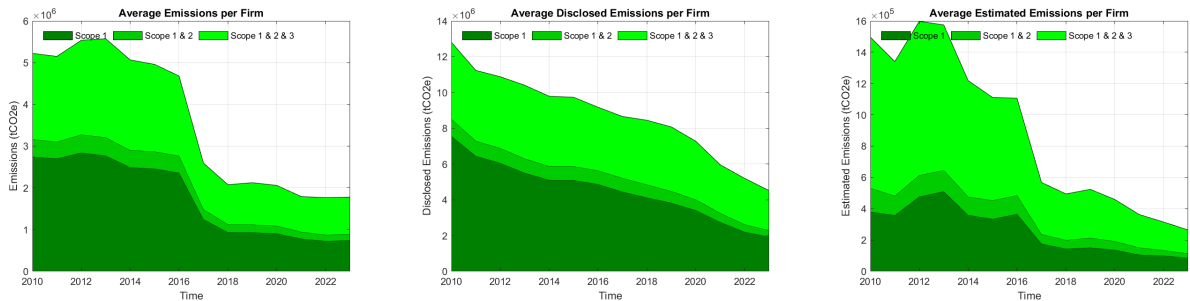


Figure 2: Total Emissions over the sample period 2010 to 2023.

**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 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).

---

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

## 5.2 Sample selection results

We begin with the sample selection equation (5) and obtain the estimates of $\beta$ by lasso probit in (9) that allows us to handle high-dimensional $Z_i$ and undertake variable selection in the parametrised propensity score $Z_i\beta$ for which $\beta \in \mathbb{R}^p$. 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).
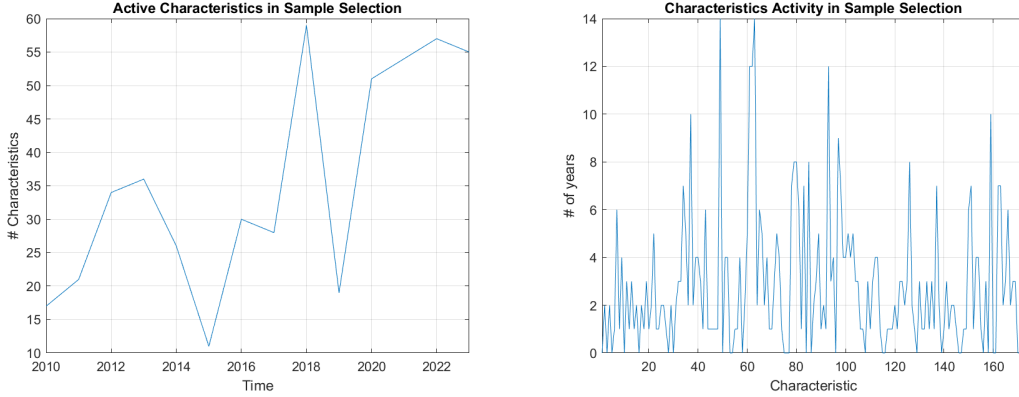


Figure 3: Active Characteristics in Sample Selection

While the number of characteristics needed to predict disclosure fluctuates, it grows substantially across our sample period and doubled between the beginning and the end of the sample period. Part of the story lies in the small and medium size of added firms. As we can see in the right panel, only few characteristics are systematically selected over the 14 years sample period, while more than half of the characteristics show at most half of the time. Three characteristics frequently play a pivotal role in sample selection: "Size" (measured by log of market capitalisation), "Age" as measured by the number of months the company is present in CRSP files, and two months lagged trading volume in Dollars ("DolVol"). These variables belong to three key variables related to firm quality: size, information quality ("Age"), and liquidity ("Dolvol"). It implies that firms with higher market equity values and good informational environment are inclined to disclose. It is in line with findings in the literature, including Flammer et al. (2021), Gibbons (2024), Krueger et al. (2024), and Gehricke et al. (2025). The relatively higher disclosure propensity observed among larger firms may reflect their greater resources for estimating aggregated carbon emissions - whether through direct measurement or indirect approximation - as well as heightened regulatory scrutiny and stronger incentives to demonstrate corporate social responsibility. The constellation of $\hat{\beta}$ estimates can be found in Figure OA.2. We also find that profitability as measured by analyst earnings per share ("FEPS"), growth as measured by sales growth and corresponding firm rank amongst peers ("MeanRankRevGrowth"), and pension funding ("FR"), which can be related to employees well-being, are important for the

26

disclosure propensity as they show up 12 out of 14 years. Only 31 out 173 characteristics are irrelevant over our sample period lending support to the need for a high-dimensional approach to understand firm disclosure decision.

**Coefficient of sample selection bias term** To correct for sample selection bias from the sample selection equation, we proceed with estimating $\theta$ for each calendar year using Neyman orthogonal score defined in (15). It turns out that the M-estimator $\theta$ satisfies the moment condition specified in (17). 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 $\theta$, and a thirty-run gives us the distribution of $\theta$ 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 $\theta$ distribution is suggested by Chernozhukov et al. (2018) in their Definition 3.3.
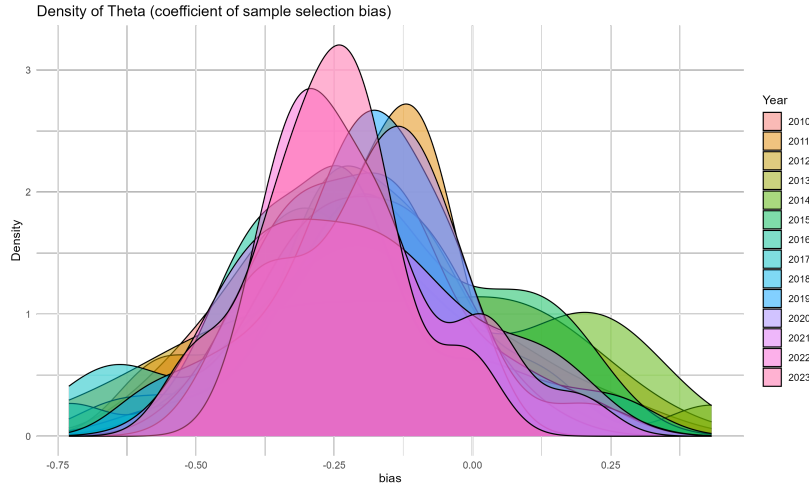


Figure 4: Distribution of $\theta$ across year 2010-2023 for Scope 1

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.

| | $\hat{\theta}$ | $\sigma_\psi$ | $\frac{\hat{\theta}}{\sigma_\psi}$ | 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 (19) 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.

## 5.3   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.
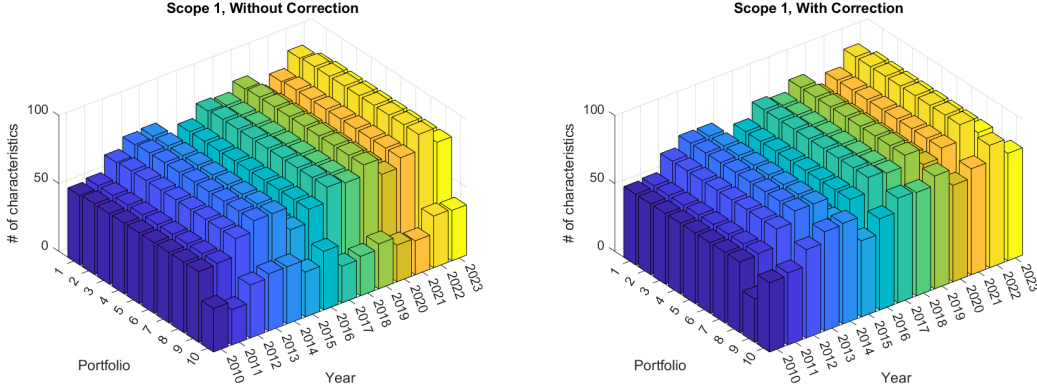
Figure 5: Active Characteristics Portfolios in Variable Selection, Scope 1

We document an apparent increase in the relevance of firm characteristics in variable selection over time, partially driven by the greater heterogeneity introduced by the sample expansion in 2016. This pattern is consistent across portfolios and holds both with and without the application of sample bias correction. A critical role of the bias correction is to restore the relevance of the extreme top decile of characteristic-sorted portfolios. While these portfolios exhibit limited explanatory power without correction, they become as relevant as other portfolios once correction is applied. This result provides an additional reason for caution in interpreting reported findings in the literature. It is well established that extreme portfolios (top and bottom decile portfolios in our case) constitute the fundamental building blocks of the 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 the short or the long legs of some factors is problematic.

The most striking result is the absence of any role for portfolios sorted on size, despite size being pivotal in sample selection as revealed in the previous section. Although sample inclusion is strongly driven by size - potentially raising identification concerns - the adaptive weights impose exclusion restrictions that mitigate this issue. After bias correction, only one additional characteristic remains inactive, i.e., not selected, 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.[11] Overall, less than 10% of the portfolios comprise the list of mismatched characteristics selected in **M1** and **M2**. The mismatched number increases from 86 in the beginning

---

[11]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"), ("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.

of the period to 146 in 2022, to be compared to 1730 portfolios each year.

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).[12] Figure OA.3 presents the estimated coefficients of the selected characteristics in **M1** and **M2**. The findings for Scope 2 and Scope 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. The variation in selected characteristics across scopes indicates that each captures distinct economic dimensions of firm emissions.

## 5.4   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** .

**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

---

[12]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").

| 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)

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

$$\hat{Y}_i = \mathsf{E}\big[Y_i\big|X_i, Z_i, D_i = 0\big] \qquad = \mathbf{k}_i\hat{b} + \mathsf{E}\big[\epsilon_i\big|X_i, Z_i, D_i = 0\big] = \mathbf{k}_i\hat{b} - \hat{\theta}\frac{\phi(-Z_i\hat{\beta})}{\Phi(-Z_i\hat{\beta})}. \qquad (28)$$

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. The idea in the estimation of the undisclosed emissions through (28) is similar to the use of a parametric model to model 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 Bia et al. (2024) and Huber and Melly (2015) for further discussion. 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.

## 5.5 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 $\hat{\theta}\frac{\phi(-Z_i\hat{\beta})}{\Phi(-Z_i\hat{\beta})}$ 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 %), 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
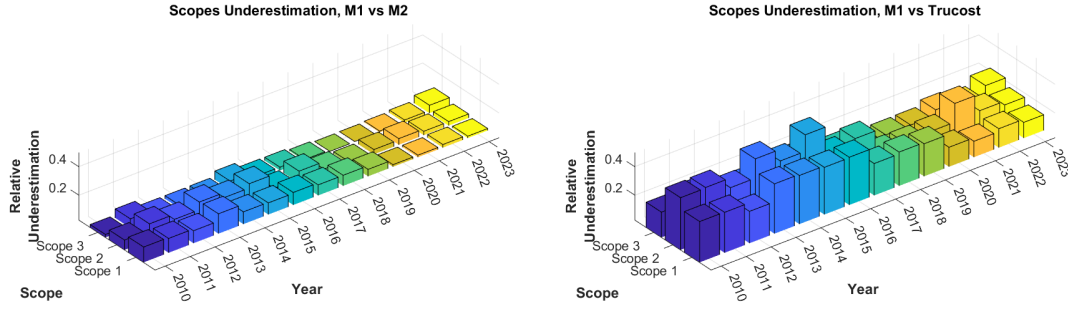
Figure 6: Relative Underestimation in Scope Emissions

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.[13] Using the average 2023 euro-to-dollar exchange rate of 0.924, this corresponds to a carbon price of $2.54 per tCO2e.[14] 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 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

---

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

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

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.
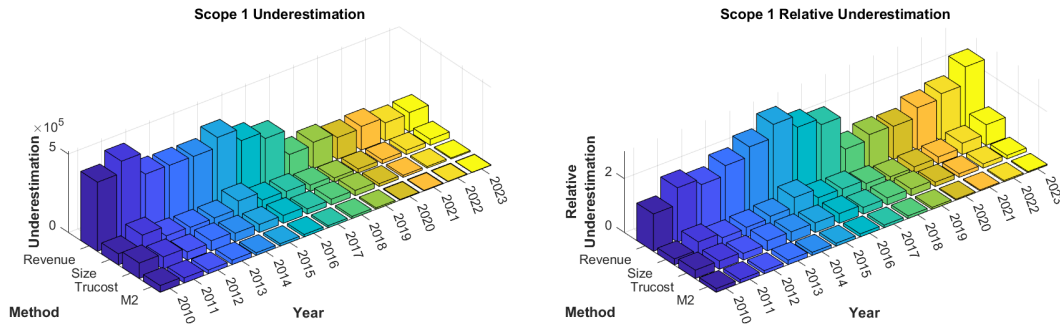


Figure 7: Underestimation in Scope 1 Emissions by Naive imputation.

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.

## 5.6   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).[15] 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.[16] 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.[17]

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.

---

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

[16]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.

[17]See Tables OA.11 - OA.13 (Scope 2) and OA.14 - OA.16 (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.

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.

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.

| | 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 understatement 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 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%).

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

|  | 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

among low emitters, consistent with where classification gaps are largest and where carbon premium tests are most sensitive.

# 6 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

Aitchison, J. and Silvey, S. (1958). Maximum-likelihood estimation of parameters subject to restraints, *The Annals of Mathematical Statistics* pp. 813–828.

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.

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

Bia, M., Huber, M. and Lafférs, 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.

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.

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.

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

Fu, W. and Knight, K. (2000). Asymptotics for lasso-type estimators, *The Annals of Statistics* **28**(5): 1356–1378.

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.

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.

Melino, A. (1982). Testing for sample selection bias, *The Review of Economic Studies* **49**(1): 151–153.

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., 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.

Silvey, S. D. (1959). The lagrangian multiplier test, *The Annals of Mathematical Statistics* **30**(2): 389–407.

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.

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.

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 $-\mathbf{k}(j)^\top(Y - \mathbf{k}\hat{b}) + \lambda w_j \frac{\mathsf{K}_j \hat{b}_j}{\|\hat{b}_j\|_{\mathsf{K}_j}} = \mathbf{0}, \forall \hat{b}_j \neq \mathbf{0}$, and $\left\| -\mathbf{k}(j)^\top(Y - \mathbf{k}\hat{b}) \right\| \leq \lambda w_j \|\mathsf{K}_j\|, \forall \hat{b}_j = \mathbf{0}$.

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 $\hat{b}$

$$\hat{b}_1 = (\mathbf{k}_1^\top \mathbf{k}_1)^{-1}(\mathbf{k}_1^\top Y - \lambda_n \mathbf{s}_1) = b_{10} + \frac{1}{n}\Sigma_{\mathbf{k}_1}^{-1}(\mathbf{k}_1^\top \boldsymbol{\epsilon} - \lambda_n \mathbf{s}_1). \tag{29}$$

If $\hat{b}_1 =_s b_{10}$, then KKT condition holds for $\hat{b} = (\hat{b}_1^\top, \mathbf{0}^\top)^\top$. 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

$$\hat{b} =_s b_0, \quad \text{if} \begin{cases} \hat{b}_1 =_s b_{10}, \\ \left\| \mathbf{k}(j)^\top(Y - \mathbf{k}_1 \hat{b}_1) \right\| < \lambda_n w_j \|\mathsf{K}_j\|, \quad \forall j \neq \mathsf{A}_b. \end{cases} \tag{30}$$

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 (29), it follows that $Y - \mathbf{k}_1 \hat{b}_1 = \boldsymbol{\epsilon} - \mathbf{k}_1(\hat{b}_1 - b_{10}) = H_n \boldsymbol{\epsilon} + \mathbf{k}_1 \Sigma_{\mathbf{k}_1}^{-1} \mathbf{s}_1 \lambda_n / n$, and by (30), we get

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

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 $\eta$ and $\eta < 1$, and $\mathbf{k}(j)^\top(H_n \boldsymbol{\epsilon})$ converges to normality asymptotically. With condition (5) and $\lambda_n w_j \to \infty$, it completes the proof since $\lim_{n \to \infty} \mathsf{P}[j \in \mathsf{A}_b] = \mathsf{P}\left[ \left\| \mathbf{k}(j)^\top(H_n \boldsymbol{\epsilon} + \mathbf{k}_1 \Sigma_{\mathbf{k}_1}^{-1} \mathbf{s}_1 \lambda_n / n) \right\| = \lambda_n w_j \frac{\mathsf{K}_j \operatorname{sgn}(\hat{b}_j)}{\|\hat{b}_j\|_{\mathsf{K}_j}} \right] \to 0$, provided that with a high probability $\lim_{n \to \infty} \mathsf{P}[j \notin \mathsf{A}_b] \to 1$.

$\square$

*Proof of Theorem 3.* We present a limiting distribution with selection bias correction. Let $b = b_0 + \frac{\boldsymbol{u}}{\sqrt{n}}$ where $\boldsymbol{u} \in \mathbb{R}^{JL \times 1}$ and $\boldsymbol{u}_j \in \mathbb{R}^{L \times 1}$. $L_n(\boldsymbol{u}) = \frac{1}{2} \left\| \frac{1}{\sqrt{n}} \mathbf{k} \boldsymbol{\mu} + \boldsymbol{\epsilon} \right\|^2 + \frac{\lambda_n}{2} \sum_{j=1}^J w_j (b_{j0} + \frac{\boldsymbol{u}_j}{\sqrt{n}})^\top \mathsf{K}_j (b_{j0} + \frac{\boldsymbol{u}_j}{\sqrt{n}})$. 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 $\hat{\boldsymbol{u}}_n = \sqrt{n}(\hat{b} - b_0)$. 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 $V_1^{(n)}(\boldsymbol{u}) = \frac{1}{2} \boldsymbol{u}^\top \Sigma_{\mathbf{k}} \boldsymbol{u} - \frac{1}{\sqrt{n}} \boldsymbol{u}^\top \mathbf{k}^\top \boldsymbol{\epsilon}$ and

$V_2^{(n)}(\boldsymbol{u}) = \frac{\lambda_n}{2} \sum_{j=1}^{J} w_j \mathsf{K}_j \left[ \left( b_{j0} + \frac{\boldsymbol{u}_j}{\sqrt{n}} \right)^\top \left( b_{j0} + \frac{\boldsymbol{u}_j}{\sqrt{n}} \right) - b_{j0}^\top b_{j0} \right]$. We get $V_1^{(n)}(\boldsymbol{u}) = \frac{1}{2} \boldsymbol{u}^\top \Sigma_{\mathbf{k}} \boldsymbol{u} - \frac{1}{\sqrt{n}} \boldsymbol{u}^\top \mathbf{k}^\top (\hat{\Gamma} +$

$\varepsilon) = \frac{1}{2} \boldsymbol{u}^\top \Sigma_{\mathbf{k}} \boldsymbol{u} - \boldsymbol{u}^\top W - \boldsymbol{u}^\top M$, where $W := \frac{1}{\sqrt{n}} \mathbf{k}^\top \varepsilon \xrightarrow{d} \mathcal{N}(0, \sigma^2 \Sigma_{\mathbf{k}_1})$ under condition 6 and $M :=$

$\frac{1}{\sqrt{n}} \mathbf{k}^\top \hat{\Gamma} \xrightarrow{d} \mathcal{N}\big( 0, \big( \Gamma_\Theta' \Sigma_\Theta (\Gamma_\Theta')^\top \big) \Sigma_{\mathbf{k}_1} \big)$. The term $\epsilon = \hat{\Gamma} + \mathsf{v}$ is contaminated by the selection bias, and thus $\boldsymbol{u}^\top M$ is an induced estimation error from that bias, which impacts the consistency and efficiency of limiting distribution. If $\theta = 0$, the proof boils down to the conventional lasso-based asymptotic analysis. Now, for $b_{j0} \neq 0$, we have $\left[ \left( b_{j0} + \frac{\boldsymbol{u}_j}{\sqrt{n}} \right)^\top \left( b_{j0} + \frac{\boldsymbol{u}_j}{\sqrt{n}} \right) - b_{j0}^\top b_{j0} \right] = \|\boldsymbol{u}_j\|$. 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 $V_2^{(n)}(\boldsymbol{u}) \xrightarrow{p} 0$. Denoting $\boldsymbol{u}_1 = (\boldsymbol{u}_j)_{j \in \mathsf{A}_b}$ and combining $V_1^{(n)}(\boldsymbol{u})$ and $V_2^{(n)}(\boldsymbol{u})$, we obtain $V^{(n)}(\boldsymbol{u}_1) \xrightarrow{d} \frac{1}{2} \boldsymbol{u}_1^\top \Sigma_{\mathbf{k}_1} \boldsymbol{u}_1 - \boldsymbol{u}_1^\top W - \boldsymbol{u}_1^\top M$. Besides, $V^{(n)}(\boldsymbol{u})$ is convex and there exists a global minimum to satisfy $\boldsymbol{u}_1^\top \Sigma_{\mathbf{k}_1} - W - M = 0$. As a result, $\hat{\boldsymbol{u}}_1 = (W + M) \Sigma_{\mathbf{k}_1}^{-1}$. Following the epiconvergence results of Fu and Knight (2000), we conclude to the asymptotic normality result:

$\sqrt{n}(\hat{b} - b_0) \xrightarrow{d} \mathcal{N}\left( \mathbf{0}, \big( \sigma^2 + \Gamma_\Theta' \Sigma_\Theta (\Gamma_\Theta')^\top \big) \Sigma_{\mathbf{k}_1}^{-1} \right)$. $\qquad\qquad \square$

# Online Appendix

## Green Silence: Double Machine Learning Carbon Emissions Under Sample Selection Bias

Cathy Yi-Hsuan Chen[1], Abraham Lioui[2], Olivier Scaillet[3]

[1]Adam Smith Business School, University of Glasgow
[2]EDHEC Business School
[3]Université de Genève and Swiss Finance Institute

September 26, 2025

# OA.1 Additional tables and figures

| | $\hat{\theta}$ | $\sigma_\psi$ | $\frac{\hat{\theta}}{\sigma_\psi}$ | p-value $\frac{\hat{\theta}}{\sigma_\psi}$ | $s_n$ | p-value $s_n$ |
|---|---|---|---|---|---|---|
| 2010 | -0.083 | 0.031 | -2.638 | 0.008 | 5.758 | 0.016 |
| 2011 | -0.193 | 0.038 | -5.112 | 0.000 | 7.155 | 0.007 |
| 2012 | -0.148 | 0.032 | -4.648 | 0.000 | 8.477 | 0.004 |
| 2013 | -0.082 | 0.028 | -2.963 | 0.003 | 6.392 | 0.011 |
| 2014 | -0.159 | 0.032 | -4.902 | 0.000 | 7.500 | 0.006 |
| 2015 | -0.226 | 0.023 | -9.797 | 0.000 | 5.813 | 0.016 |
| 2016 | -0.070 | 0.032 | -2.213 | 0.027 | 8.325 | 0.004 |
| 2017 | -0.131 | 0.026 | -4.963 | 0.000 | 6.067 | 0.014 |
| 2018 | -0.172 | 0.022 | -7.941 | 0.000 | 4.669 | 0.031 |
| 2019 | -0.111 | 0.022 | -5.091 | 0.000 | 4.661 | 0.031 |
| 2020 | -0.146 | 0.018 | -7.990 | 0.000 | 4.976 | 0.026 |
| 2021 | -0.164 | 0.016 | -10.015 | 0.000 | 4.615 | 0.032 |
| 2022 | -0.196 | 0.016 | -12.290 | 0.000 | 5.111 | 0.024 |
| 2023 | -0.161 | 0.017 | -9.492 | 0.000 | 5.630 | 0.018 |

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

| | $\hat{\theta}$ | $\sigma_\psi$ | $\frac{\hat{\theta}}{\sigma_\psi}$ | p-value $\frac{\hat{\theta}}{\sigma_\psi}$ | $s_n$ | p-value $s_n$ |
|---|---|---|---|---|---|---|
| 2010 | -0.064 | 0.019 | -3.289 | 0.001 | 3.678 | 0.055 |
| 2011 | -0.162 | 0.020 | -8.221 | 0.000 | 4.825 | 0.028 |
| 2012 | -0.112 | 0.016 | -7.136 | 0.000 | 4.792 | 0.029 |
| 2013 | -0.080 | 0.026 | -3.017 | 0.003 | 7.919 | 0.005 |
| 2014 | -0.094 | 0.016 | -5.800 | 0.000 | 4.851 | 0.028 |
| 2015 | -0.159 | 0.017 | -9.186 | 0.000 | 6.167 | 0.013 |
| 2016 | -0.123 | 0.015 | -8.261 | 0.000 | 5.840 | 0.016 |
| 2017 | -0.113 | 0.017 | -6.668 | 0.000 | 6.348 | 0.012 |
| 2018 | -0.129 | 0.014 | -9.086 | 0.000 | 6.160 | 0.013 |
| 2019 | -0.108 | 0.015 | -7.280 | 0.000 | 6.626 | 0.010 |
| 2020 | -0.156 | 0.015 | -10.127 | 0.000 | 5.915 | 0.015 |
| 2021 | -0.112 | 0.014 | -8.036 | 0.000 | 7.875 | 0.005 |
| 2022 | -0.139 | 0.016 | -8.573 | 0.000 | 9.398 | 0.002 |
| 2023 | -0.183 | 0.018 | -10.225 | 0.000 | 14.818 | 0.000 |

Table OA.2: Significance of $\hat{\theta}$ and score test for Scope 3

|      | $\hat{\theta}$ | $\sigma_\psi$ | $\frac{\hat{\theta}}{\sigma_\psi}$ | p-value $\frac{\hat{\theta}}{\sigma_\psi}$ | $s_n$ | p-value $s_n$ |
|------|--------|--------|--------|--------|--------|--------|
| 2010 | -0.054 | 0.032 | -1.664 | 0.096 | 5.471 | 0.019 |
| 2011 | -0.126 | 0.035 | -3.611 | 0.000 | 9.168 | 0.002 |
| 2012 | -0.108 | 0.024 | -4.454 | 0.000 | 7.632 | 0.006 |
| 2013 | -0.039 | 0.023 | -1.701 | 0.089 | 7.792 | 0.005 |
| 2014 | 0.007 | 0.020 | 0.344 | 1.269 | 9.725 | 0.002 |
| 2015 | -0.107 | 0.025 | -4.267 | 0.000 | 9.423 | 0.002 |
| 2016 | -0.036 | 0.021 | -1.680 | 0.093 | 8.318 | 0.004 |
| 2017 | -0.028 | 0.019 | -1.443 | 0.149 | 7.461 | 0.006 |
| 2018 | -0.046 | 0.022 | -2.139 | 0.032 | 7.251 | 0.007 |
| 2019 | -0.031 | 0.014 | -2.190 | 0.029 | 5.438 | 0.020 |
| 2020 | -0.078 | 0.014 | -5.534 | 0.000 | 6.301 | 0.012 |
| 2021 | -0.054 | 0.018 | -3.030 | 0.002 | 7.149 | 0.008 |
| 2022 | -0.145 | 0.017 | -8.337 | 0.000 | 8.043 | 0.005 |
| 2023 | -0.105 | 0.020 | -5.140 | 0.000 | 10.078 | 0.002 |

Table OA.3: Significance of $\hat{\theta}$ and score test without mitigating impact of regularisation bias for Scope 1

| Year | Ton of CO2e | Carbon price ($) | A330 flight |
|------|--------|--------|--------|
| 2010 | 13 983 750 | 35 518 725 | 93 225 |
| 2011 | 14 213 031 | 36 101 098 | 94 754 |
| 2012 | 9 943 236 | 25 255 819 | 66 288 |
| 2013 | 8 831 187 | 22 431 214 | 58 875 |
| 2014 | 5 335 667 | 13 552 594 | 35 571 |
| 2015 | 9 570 923 | 24 310 144 | 63 806 |
| 2016 | 6 168 000 | 15 666 720 | 41 120 |
| 2017 | 6 422 712 | 16 313 688 | 42 818 |
| 2018 | 17 535 256 | 44 539 550 | 116 902 |
| 2019 | 19 270 978 | 48 948 284 | 128 473 |
| 2020 | 9 982 329 | 25 355 115 | 66 549 |
| 2021 | 5 279 789 | 13 410 664 | 35 199 |
| 2022 | 4 566 150 | 11 598 021 | 30 441 |
| 2023 | 2 268 535 | 5 762 078 | 15 124 |
| Total | 133 371 543 | 338 763 719 | 889 145 |

Table OA.4: Underestimation of **M2** in tCO2e without DML

The first column is the estimated underestimation of emitted tCO2e. The second column is to have first column multiplied by carbon price in the US ETS and carbon tax, amounting to $2.54 per tCO2e. The third column translates the first column tino the number of trip operated by A330 aircraft from London (GB), LHR to: New York (US), JFK, single trip, ca. 5,550 km, amounting to 150 tCO2e.

| Year | MSE of **M1** | MSE of **M2** | Relative error |
|------|------|------|------|
| 2010 | 1.076 | 2.626 | 0.410 |
| 2011 | 1.432 | 1.881 | 0.761 |
| 2012 | 1.043 | 1.503 | 0.693 |
| 2013 | 1.935 | 5.199 | 0.372 |
| 2014 | 1.452 | 2.862 | 0.507 |
| 2015 | 2.381 | 2.482 | 0.959 |
| 2016 | 1.667 | 2.044 | 0.815 |
| 2017 | 2.091 | 2.534 | 0.825 |
| 2018 | 1.940 | 2.281 | 0.851 |
| 2019 | 1.599 | 1.707 | 0.937 |
| 2020 | 1.807 | 2.905 | 0.622 |
| 2021 | 1.647 | 1.931 | 0.853 |
| 2022 | 1.522 | 2.166 | 0.703 |
| 2023 | 1.143 | 1.367 | 0.836 |

Table OA.5: Prediction performance (Scope 2)

The third column is the relative MSE, a ratio of column 1 and column 2.

| Year | MSE of **M1** | MSE of **M2** | Relative error |
|------|------|------|------|
| 2010 | 0.445 | 0.783 | 0.569 |
| 2011 | 0.572 | 0.652 | 0.877 |
| 2012 | 0.833 | 0.901 | 0.925 |
| 2013 | 0.607 | 0.653 | 0.930 |
| 2014 | 0.885 | 0.961 | 0.920 |
| 2015 | 0.578 | 0.637 | 0.907 |
| 2016 | 0.732 | 0.780 | 0.938 |
| 2017 | 0.662 | 0.728 | 0.909 |
| 2018 | 0.833 | 0.994 | 0.838 |
| 2019 | 0.828 | 0.854 | 0.970 |
| 2020 | 0.664 | 0.682 | 0.973 |
| 2021 | 0.628 | 0.678 | 0.926 |
| 2022 | 0.826 | 0.904 | 0.913 |
| 2023 | 1.236 | 1.349 | 0.916 |

Table OA.6: Prediction performance (Scope 3)

The third column is the relative MSE, a ratio of column 1 and column 2.

| Year | median (tCO2e) | aggregation (tCO2e) | Carbon price ($) | A330 flight |
|------|------|------|------|------|
| | | Underestimation of **M2** | | |
| 2010 | 41977 | 24514421 | 62266629 | 163429 |
| 2011 | 32103 | 17495903 | 44439595 | 116639 |
| 2012 | 23488 | 11462330 | 29114319 | 76416 |
| 2013 | 28347 | 13323021 | 33840474 | 88820 |
| 2014 | 16282 | 7310741 | 18569282 | 48738 |
| 2015 | 15330 | 7986919 | 20286775 | 53246 |
| 2016 | 15581 | 7681483 | 19510966 | 51210 |
| 2017 | 13200 | 8804215 | 22362705 | 58695 |
| 2018 | 13763 | 25943442 | 65896344 | 172956 |
| 2019 | 9748 | 18404425 | 46747240 | 122696 |
| 2020 | 4465 | 8183527 | 20786159 | 54557 |
| 2021 | 3740 | 6806420 | 17288306 | 45376 |
| 2022 | 3045 | 6047800 | 15361412 | 40319 |
| 2023 | 1947 | 3657797 | 9290804 | 24385 |
| Total | 223015 | 167622445 | 425761010 | 1117483 |
| | | Underestimation of Trucost | | |
| 2010 | 98928 | 57774234 | 146746553 | 385162 |
| 2011 | 78996 | 43053067 | 109354790 | 287020 |
| 2012 | 51962 | 25357583 | 64408260 | 169051 |
| 2013 | 64663 | 30391413 | 77194189 | 202609 |
| 2014 | 53936 | 24217067 | 61511351 | 161447 |
| 2015 | 51326 | 26741100 | 67922395 | 178274 |
| 2016 | 49835 | 24568629 | 62404319 | 163791 |
| 2017 | 33223 | 22159449 | 56285000 | 147730 |
| 2018 | 36445 | 68698410 | 174493960 | 457989 |
| 2019 | 33597 | 63431194 | 161115233 | 422875 |
| 2020 | 18336 | 33610077 | 85369596 | 224067 |
| 2021 | 13089 | 23821485 | 60506573 | 158810 |
| 2022 | 10343 | 20540901 | 52173889 | 136939 |
| 2023 | 6840 | 12851731 | 32643398 | 85678 |
| Total | 601518 | 477216340 | 1212129504 | 3181442 |

Table OA.7: Underestimation in tCO2e and pecuniary measures (Scope1)

The first column is the estimated underestimation of emitted tCO2e. The second column is to have first column multiplied by carbon price in the US ETS and carbon tax, amounting to $2.54 per tCO2e. The third column translates the first column into the number of trip operated by A330 aircraft from London (LHR) to New York (JFK), single trip, ca. 5,550 km, amounting to 150 tCO2e.

| Year | median (tCO2e) | aggregation (tCO2e) | Carbon price ($) | A330 flight |
|---|---|---|---|---|
| | | Underestimation of **M2** | | |
| 2010 | 27338 | 15965250 | 40551736 | 106435 |
| 2011 | 38091 | 20759361 | 52728777 | 138396 |
| 2012 | 28262 | 13791728 | 35030990 | 91945 |
| 2013 | 40296 | 18939315 | 48105861 | 126262 |
| 2014 | 28760 | 12913380 | 32799986 | 86089 |
| 2015 | 35813 | 18658645 | 47392959 | 124391 |
| 2016 | 12360 | 6093689 | 15477970 | 40625 |
| 2017 | 21300 | 14207004 | 36085789 | 94713 |
| 2018 | 10181 | 19191484 | 48746369 | 127943 |
| 2019 | 6990 | 13196269 | 33518522 | 87975 |
| 2020 | 8981 | 16463026 | 41816086 | 109754 |
| 2021 | 7462 | 13580253 | 34493843 | 90535 |
| 2022 | 4048 | 8040073 | 20421786 | 53600 |
| 2023 | 3080 | 5787979 | 14701467 | 38587 |
| Total | 272963 | 197587457 | 501872141 | 1317250 |
| | | Underestimation of Trucost | | |
| 2010 | 114161 | 66670024 | 169341860 | 444467 |
| 2011 | 107260 | 58456519 | 148479558 | 389710 |
| 2012 | 88458 | 43167713 | 109645991 | 287785 |
| 2013 | 117031 | 55004749 | 139712062 | 366698 |
| 2014 | 88699 | 39826022 | 101158095 | 265507 |
| 2015 | 115833 | 60349148 | 153286836 | 402328 |
| 2016 | 66366 | 32718432 | 83104816 | 218123 |
| 2017 | 66397 | 44286548 | 112487831 | 295244 |
| 2018 | 32028 | 60372461 | 153346051 | 402483 |
| 2019 | 29991 | 56622360 | 143820794 | 377482 |
| 2020 | 28500 | 52239716 | 132688878 | 348265 |
| 2021 | 34642 | 63048245 | 160142542 | 420322 |
| 2022 | 15328 | 30440664 | 77319286 | 202938 |
| 2023 | 9871 | 18547476 | 47110588 | 123650 |
| Total | 914564 | 681750075 | 1731645190 | 4545000 |

Table OA.8: Underestimation in tCO2e and pecuniary measures (Scope2)

The first column is the estimated underestimation of emitted tCO2e. The second column is to have first column multiplied by carbon price in the US ETS and carbon tax, amounting to $2.54 per tCO2e. The third column translates the first column into the number of trip operated by A330 aircraft from London (LHR) to New York (JFK), single trip, ca. 5,550 km, amounting to 150 tCO2e.

| Year | median (tCO2e) | aggregation (tCO2e) | Carbon price ($) | A330 flight |
|---|---|---|---|---|
| | | Underestimation of **M2** | | |
| 2010 | 70615 | 41239063 | 104747221 | 274927 |
| 2011 | 131158 | 71481293 | 181562484 | 476542 |
| 2012 | 67862 | 33116863 | 84116833 | 220779 |
| 2013 | 63695 | 29936843 | 76039581 | 199579 |
| 2014 | 64378 | 28905751 | 73420608 | 192705 |
| 2015 | 114140 | 59466822 | 151045727 | 396445 |
| 2016 | 61723 | 30429252 | 77290299 | 202862 |
| 2017 | 57635 | 38442409 | 97643719 | 256283 |
| 2018 | 40008 | 75415331 | 191554940 | 502769 |
| 2019 | 26220 | 49503875 | 125739843 | 330026 |
| 2020 | 25546 | 46825024 | 118935560 | 312167 |
| 2021 | 24141 | 43936446 | 111598572 | 292910 |
| 2022 | 27456 | 54527035 | 138498670 | 363514 |
| 2023 | 38719 | 72752577 | 184791545 | 485017 |
| Total | 813296 | 675978583 | 1716985601 | 4506524 |
| | | Underestimation of Trucost | | |
| 2010 | 315008 | 183964797 | 467270585 | 1226432 |
| 2011 | 362437 | 197528316 | 501721924 | 1316855 |
| 2012 | 242702 | 118438725 | 300834362 | 789592 |
| 2013 | 253915 | 119339895 | 303123332 | 795599 |
| 2014 | 261274 | 117311886 | 297972190 | 782079 |
| 2015 | 300161 | 156384047 | 397215479 | 1042560 |
| 2016 | 255064 | 125746602 | 319396369 | 838311 |
| 2017 | 217129 | 144824988 | 367855470 | 965500 |
| 2018 | 137773 | 259702638 | 659644700 | 1731351 |
| 2019 | 148722 | 280787886 | 713201230 | 1871919 |
| 2020 | 136384 | 249991820 | 634979223 | 1666612 |
| 2021 | 132416 | 240996251 | 612130478 | 1606642 |
| 2022 | 97794 | 194219689 | 493318009 | 1294798 |
| 2023 | 105642 | 198501421 | 504193610 | 1323343 |
| Total | 2966422 | 2587738961 | 6572856961 | 17251593 |

Table OA.9: Underestimation in tCO2e and pecuniary measures (Scope3)

The first column is the estimated underestimation of emitted tCO2e. The second column is to have first column multiplied by carbon price in the US ETS and carbon tax, amounting to \$2.54 per tCO2e. The third column translates the first column into the number of trip operated by A330 aircraft from London (LHR) to New York (JFK), single trip, ca. 5,550 km, amounting to 150 tCO2e.

| Year | median (tCO2e) | aggregation (tCO2e) | Carbon price ($) | A330 flight |
|---|---|---|---|---|
| Underestimation (Scope1) of using `size` only | | | | |
| 2010 | 70615 | 41239063 | 104747221 | 274927 |
| 2011 | 131158 | 71481293 | 181562484 | 476542 |
| 2012 | 67862 | 33116863 | 84116833 | 220779 |
| 2013 | 63695 | 29936843 | 76039581 | 199579 |
| 2014 | 64378 | 28905751 | 73420608 | 192705 |
| 2015 | 114140 | 59466822 | 151045727 | 396445 |
| 2016 | 61723 | 30429252 | 77290299 | 202862 |
| 2017 | 57635 | 38442409 | 97643719 | 256283 |
| 2018 | 40008 | 75415331 | 191554940 | 502769 |
| 2019 | 26220 | 49503875 | 125739843 | 330026 |
| 2020 | 25546 | 46825024 | 118935560 | 312167 |
| 2021 | 24141 | 43936446 | 111598572 | 292910 |
| 2022 | 27456 | 54527035 | 138498670 | 363514 |
| 2023 | 38719 | 72752577 | 184791545 | 485017 |
| Total | 813296 | 675978583 | 1716985601 | 4506524 |
| Underestimation (Scope1) of using `revenue` only | | | | |
| 2010 | 441051 | 257573905 | 654237718 | 1717159 |
| 2011 | 516299 | 281382809 | 714712335 | 1875885 |
| 2012 | 375042 | 183020600 | 464872325 | 1220137 |
| 2013 | 372538 | 175092679 | 444735405 | 1167285 |
| 2014 | 370011 | 166134947 | 421982767 | 1107566 |
| 2015 | 433010 | 225598304 | 573019692 | 1503989 |
| 2016 | 354223 | 174631839 | 443564872 | 1164212 |
| 2017 | 308999 | 206102339 | 523499942 | 1374016 |
| 2018 | 147915 | 278819357 | 708201166 | 1858796 |
| 2019 | 167161 | 315599294 | 801622206 | 2103995 |
| 2020 | 138898 | 254599482 | 646682685 | 1697330 |
| 2021 | 150884 | 274609056 | 697507002 | 1830727 |
| 2022 | 115046 | 228480667 | 580340895 | 1523204 |
| 2023 | 122055 | 229341727 | 582527986 | 1528945 |
| Total | 4013131 | 3250987006 | 8257506996 | 21673247 |

Table OA.10: Underestimation casued by simply using size or revenue

The first column is the estimated underestimation of emitted tCO2e. The second column is to have first column multiplied by carbon price in the US ETS and carbon tax, amounting to $2.54 per tCO2e. The third column translates the first column into the number of trip operated by A330 aircraft from London (LHR) to New York (JFK), single trip, ca. 5,550 km, amounting to 150 tCO2e.

| Rank Correlation | | | | | | Jaccard Index | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **M1** | **M2** | **Vendor** | **Size** | **Revenue** | | **M2** | **Vendor** | **Size** | **Revenue** |
| M1 | 1.00 | 0.80 | 0.71 | 0.60 | 0.63 | Bottom 10% | 0.41 | 0.14 | 0.31 | 0.32 |
| M2 | 0.80 | 1.00 | 0.58 | 0.55 | 0.64 | Bottom 30% | 0.53 | 0.39 | 0.35 | 0.38 |
| Vendor | 0.71 | 0.58 | 1.00 | 0.46 | 0.47 | Bottom 50% | 0.65 | 0.57 | 0.50 | 0.53 |
| Size | 0.60 | 0.55 | 0.46 | 1.00 | 0.64 | Top 50% | 0.65 | 0.60 | 0.50 | 0.52 |
| Revenue | 0.63 | 0.64 | 0.47 | 0.64 | 1.00 | Top 30% | 0.51 | 0.58 | 0.46 | 0.48 |
| | | | | | | Top 10% | 0.40 | 0.78 | 0.78 | 0.75 |

Table OA.11: Agreement Across Carbon Emissions Estimation Methods. Scope 2

| | High Emitters | | | | | Low Emitters | | | | | High minus Low Emitters | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **M1** | **M2** | **Vendor** | **Size** | **Revenue** | **M1** | **M2** | **Vendor** | **Size** | **Revenue** | **M1** | **M2** | **Vendor** | **Size** | **Revenue** |
| Mean | 1.51 | 1.47 | 1.51 | 1.50 | 1.47 | 1.72 | 1.66 | 1.82 | 1.66 | 1.62 | -0.21 | -0.20 | -0.30 | -0.17 | -0.16 |
| $t$(Mean) | 4.54 | 4.45 | 4.60 | 4.54 | 4.41 | 4.94 | 4.75 | 4.59 | 4.71 | 4.71 | -1.63 | -1.70 | -1.36 | -1.53 | -1.40 |
| Median | 1.81 | 1.74 | 1.85 | 1.90 | 1.84 | 2.06 | 2.09 | 1.64 | 2.04 | 2.03 | -0.29 | -0.32 | -0.34 | -0.18 | -0.28 |
| Std | 4.29 | 4.27 | 4.27 | 4.27 | 4.31 | 4.51 | 4.54 | 5.14 | 4.57 | 4.46 | 1.69 | 1.49 | 2.91 | 1.40 | 1.44 |
| Skewness | -0.15 | -0.13 | -0.15 | -0.14 | -0.15 | -0.30 | -0.21 | 0.05 | -0.28 | -0.19 | 0.12 | 0.10 | -0.09 | 0.08 | 0.21 |
| Kurtosis | 3.47 | 3.40 | 3.45 | 3.41 | 3.46 | 3.95 | 3.33 | 4.07 | 3.55 | 3.48 | 3.74 | 3.12 | 3.95 | 4.00 | 3.32 |
| VaR(5%) | -6.22 | -6.55 | -6.07 | -6.10 | -6.50 | -6.28 | -6.74 | -6.41 | -6.54 | -6.47 | -3.08 | -2.41 | -4.71 | -2.58 | -2.35 |
| ES(5%) | -8.24 | -8.09 | -8.17 | -8.08 | -8.29 | -8.82 | -8.59 | -9.60 | -9.00 | -8.40 | -3.87 | -3.35 | -7.05 | -3.34 | -3.25 |

Table OA.12: Carbon Premium. Scope 2

| | High Emitters | | | | | Low Emitters | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **M1** | **M2** | **Vendor** | **Size** | **Revenue** | **M1** | **M2** | **Vendor** | **Size** | **Revenue** |
| | **Carbon Emissions (in Thousands)** | | | | | | | | | |
| Mean | 2082.00 | 2662.69 | 2037.22 | 2039.79 | 2219.40 | 29.46 | 61.08 | 7.59 | 42.46 | 74.67 |
| $t$(Mean) | 54.57 | 52.25 | 67.74 | 69.10 | 62.22 | 23.95 | 38.77 | 16.52 | 41.17 | 42.54 |
| Median | 1963.34 | 2522.89 | 1889.03 | 1917.54 | 1995.99 | 35.34 | 66.74 | 10.39 | 46.36 | 80.80 |
| Std | 494.54 | 660.58 | 389.81 | 382.61 | 462.34 | 15.94 | 20.42 | 5.95 | 13.37 | 22.75 |
| Skewness | 0.34 | 0.68 | 0.75 | 0.39 | 0.75 | -0.14 | -0.30 | -0.02 | -0.91 | -0.94 |
| Kurtosis | 1.90 | 2.49 | 2.47 | 1.80 | 2.48 | 1.49 | 2.80 | 1.15 | 2.83 | 2.90 |
| | **Characteristics** | | | | | | | | | |
| Size | 18.22 | 18.20 | 18.23 | 18.28 | 18.26 | 16.20 | 16.86 | 15.49 | 16.56 | 16.88 |
| Book-to-market | 1.63 | 1.65 | 1.62 | 1.62 | 1.63 | 1.81 | 1.67 | 1.91 | 1.76 | 1.74 |
| Profitability | 0.19 | 0.19 | 0.19 | 0.20 | 0.19 | 0.17 | 0.19 | 0.13 | 0.18 | 0.18 |
| Investment | 0.10 | 0.09 | 0.10 | 0.10 | 0.09 | 0.14 | 0.13 | 0.20 | 0.13 | 0.13 |
| Leverage | 3.60 | 3.63 | 3.59 | 3.60 | 3.56 | 3.57 | 3.72 | 4.02 | 3.65 | 3.92 |
| Revenue | 57.44 | 60.80 | 56.63 | 58.50 | 59.56 | 6.14 | 11.45 | 2.64 | 8.43 | 14.64 |

Table OA.13: Carbon Factors Emissions and Characteristics. Scope 2.

| | **Rank Correlation** | | | | | | **Jaccard Index** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **M1** | **M2** | **Vendor** | **Size** | **Revenue** | | **M2** | **Vendor** | **Size** | **Revenue** |
| M1 | 1.00 | 0.99 | 0.80 | 0.56 | 0.59 | Bottom 10% | 0.75 | 0.20 | 0.17 | 0.13 |
| M2 | 0.99 | 1.00 | 0.81 | 0.56 | 0.58 | Bottom 30% | 0.87 | 0.48 | 0.29 | 0.37 |
| Vendor | 0.80 | 0.81 | 1.00 | 0.45 | 0.47 | Bottom 50% | 0.92 | 0.65 | 0.48 | 0.48 |
| Size | 0.56 | 0.56 | 0.45 | 1.00 | 0.62 | Top 50% | 0.92 | 0.69 | 0.48 | 0.48 |
| Revenue | 0.59 | 0.58 | 0.47 | 0.62 | 1.00 | Top 30% | 0.89 | 0.66 | 0.45 | 0.47 |
| | | | | | | Top 10% | 0.91 | 0.79 | 0.76 | 0.76 |

Table OA.14: Agreement Across Carbon Emissions Estimation Methods. Scope 3

| | **High Emitters** | | | | | **Low Emitters** | | | | | **High minus Low Emitters** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **M1** | **M2** | **Vendor** | **Size** | **Revenue** | **M1** | **M2** | **Vendor** | **Size** | **Revenue** | **M1** | **M2** | **Vendor** | **Size** | **Revenue** |
| Mean | 1.47 | 1.46 | 1.49 | 1.51 | 1.50 | 1.80 | 1.77 | 1.82 | 1.74 | 1.63 | -0.33 | -0.31 | -0.33 | -0.22 | -0.13 |
| $t$(Mean) | 4.56 | 4.53 | 4.58 | 4.68 | 4.63 | 4.67 | 4.59 | 4.49 | 4.70 | 4.60 | -1.87 | -1.72 | -1.40 | -1.68 | -1.03 |
| Median | 1.92 | 1.94 | 1.86 | 1.98 | 1.84 | 2.04 | 2.00 | 1.96 | 2.12 | 1.93 | -0.52 | -0.40 | -0.21 | -0.27 | -0.18 |
| Std | 4.17 | 4.18 | 4.23 | 4.20 | 4.20 | 4.99 | 5.00 | 5.26 | 4.79 | 4.58 | 2.32 | 2.30 | 3.05 | 1.72 | 1.59 |
| Skewness | -0.20 | -0.20 | -0.16 | -0.15 | -0.13 | -0.18 | -0.15 | 0.13 | -0.19 | -0.24 | 0.19 | 0.12 | -0.06 | 0.19 | 0.42 |
| Kurtosis | 3.39 | 3.43 | 3.43 | 3.38 | 3.41 | 4.10 | 4.17 | 3.87 | 3.66 | 3.73 | 3.23 | 3.36 | 4.28 | 3.33 | 4.61 |
| VaR(5%) | -6.26 | -6.30 | -6.31 | -6.21 | -6.28 | -7.28 | -7.07 | -6.95 | -7.45 | -7.09 | -4.22 | -4.25 | -5.93 | -3.10 | -2.45 |
| ES(5%) | -8.08 | -8.12 | -8.12 | -7.99 | -7.95 | -9.71 | -9.56 | -9.71 | -9.09 | -8.93 | -5.12 | -5.09 | -7.48 | -3.78 | -3.30 |

Table OA.15: Carbon Premium. Scope 3

| | **High Emitters** | | | | | **Low Emitters** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **M1** | **M2** | **Vendor** | **Size** | **Revenue** | **M1** | **M2** | **Vendor** | **Size** | **Revenue** |
| | **Carbon Emissions (in Thousands)** | | | | | | | | | |
| Mean | 12753.23 | 12558.70 | 12539.96 | 12664.53 | 13942.33 | 148.84 | 131.71 | 54.72 | 328.30 | 432.88 |
| $t$(Mean) | 57.67 | 57.75 | 62.71 | 65.28 | 65.91 | 30.87 | 30.90 | 16.23 | 34.55 | 47.62 |
| Median | 12513.53 | 12272.91 | 12189.99 | 12888.48 | 13469.13 | 176.47 | 163.05 | 70.10 | 360.96 | 427.54 |
| Std | 2866.57 | 2818.91 | 2591.90 | 2514.75 | 2741.66 | 62.49 | 55.24 | 43.69 | 123.18 | 117.82 |
| Skewness | 0.14 | 0.11 | 0.17 | -0.00 | 0.22 | -0.16 | -0.31 | -0.06 | 0.01 | 0.71 |
| Kurtosis | 1.50 | 1.46 | 1.51 | 1.45 | 1.51 | 1.38 | 1.36 | 1.08 | 2.14 | 4.49 |
| | **Characteristics** | | | | | | | | | |
| Size | 18.23 | 18.23 | 18.22 | 18.28 | 18.27 | 15.93 | 15.88 | 15.29 | 16.46 | 16.86 |
| Book-to-market | 1.63 | 1.63 | 1.63 | 1.61 | 1.62 | 1.70 | 1.70 | 1.70 | 1.73 | 1.81 |
| Profitability | 0.20 | 0.20 | 0.19 | 0.20 | 0.19 | 0.16 | 0.16 | 0.14 | 0.17 | 0.17 |
| Investment | 0.09 | 0.09 | 0.10 | 0.11 | 0.09 | 0.16 | 0.17 | 0.21 | 0.14 | 0.14 |
| Leverage | 3.47 | 3.47 | 3.51 | 3.49 | 3.44 | 3.51 | 3.53 | 3.90 | 4.21 | 4.87 |
| Revenue | 57.89 | 57.44 | 56.98 | 59.11 | 60.91 | 3.47 | 4.20 | 1.34 | 5.82 | 13.70 |

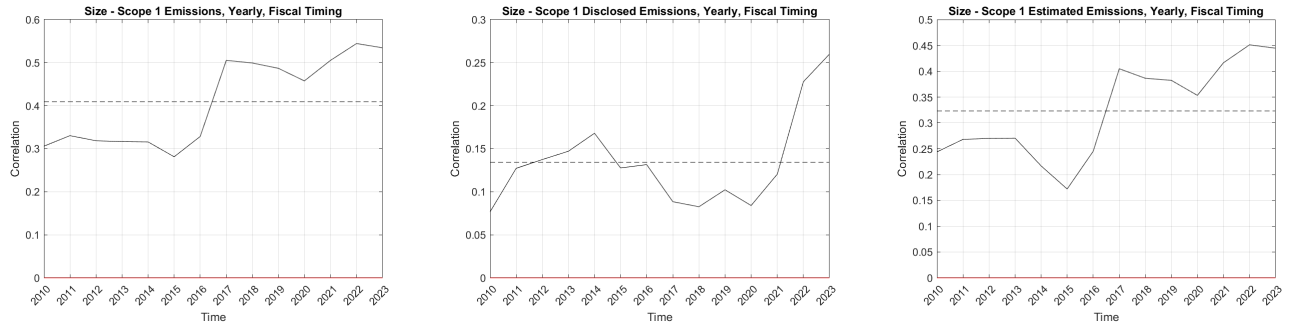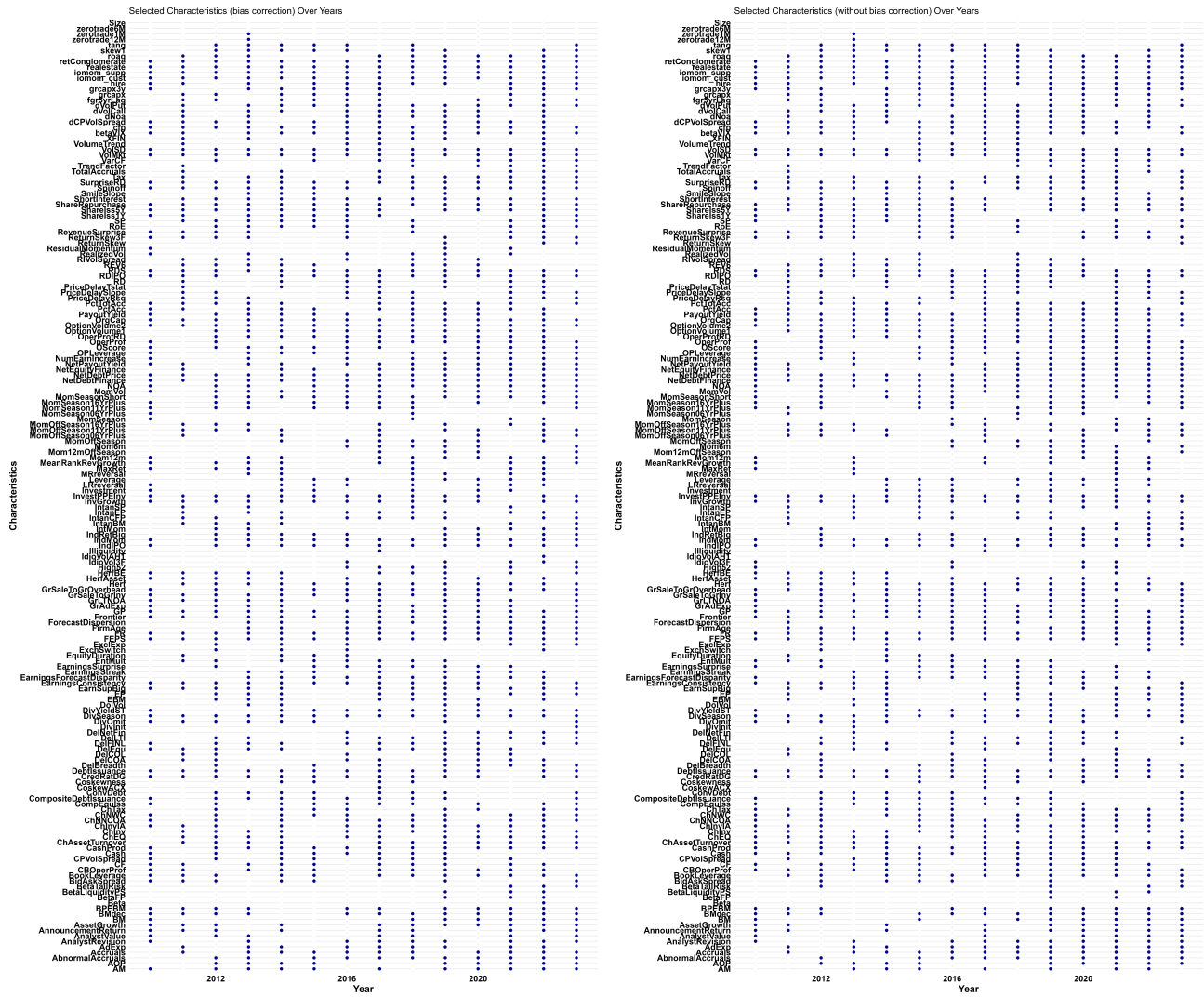Table OA.16: Carbon Factors Emissions and Characteristics. Scope 3.

Figure OA.1: Correlation with firm size (log market cap)

Figure OA.2: The non-zero $\hat{\beta}$ estimated via lasso probit in (9)
The sample period spans 2010 to 2023.

(a) with bias correction

(b) without bias correction

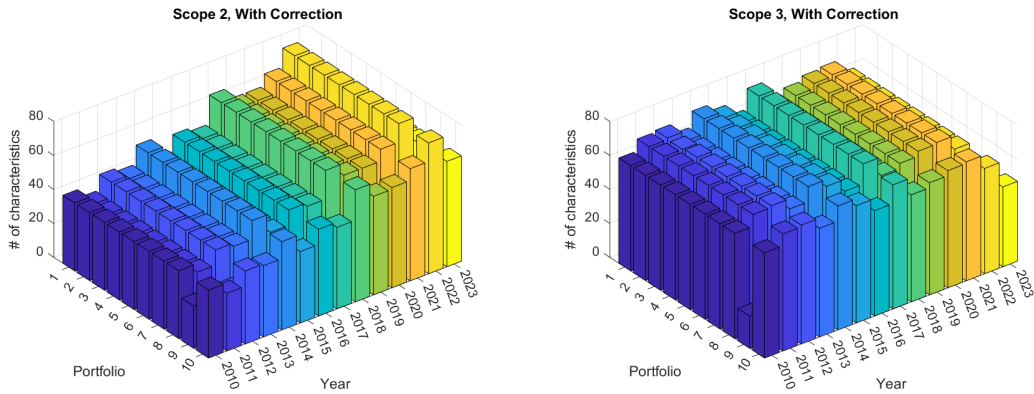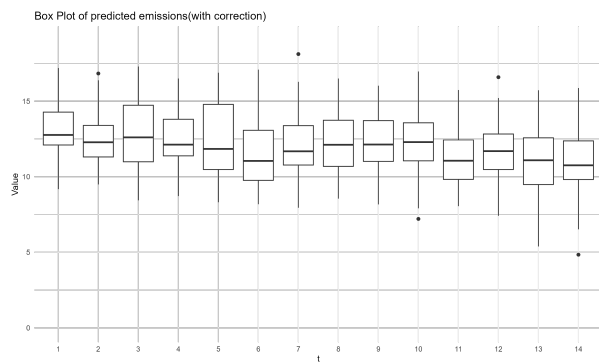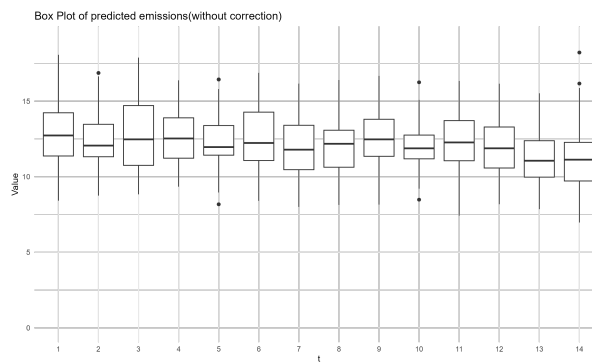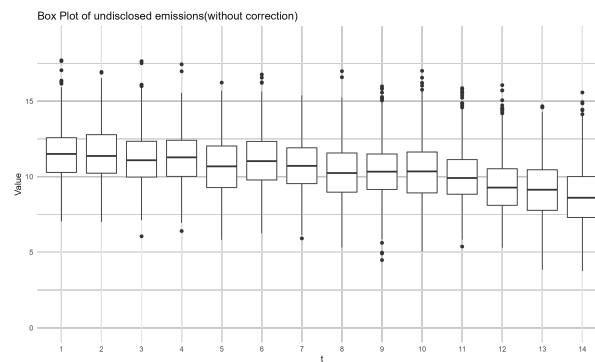Figure OA.3: Characteristics determine Scope 1 emission



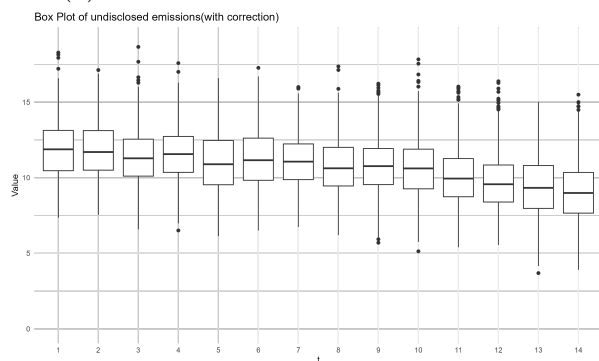Figure OA.4: Active Characteristics Portfolios in Variable Selection, Scope 2 and 3
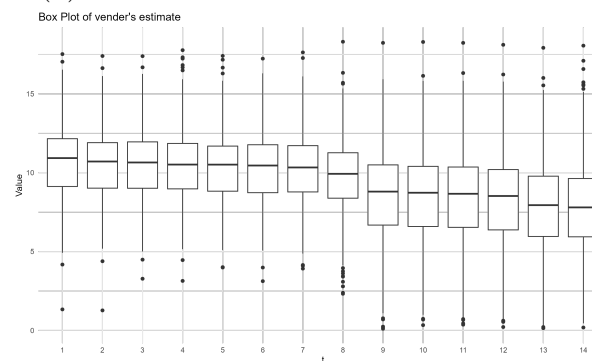
(a) Predicted disclosed with correction

(b) Predicted disclosed without correction

(c) Predicted undisclosed without correction

(d) Predicted undisclosed with correction

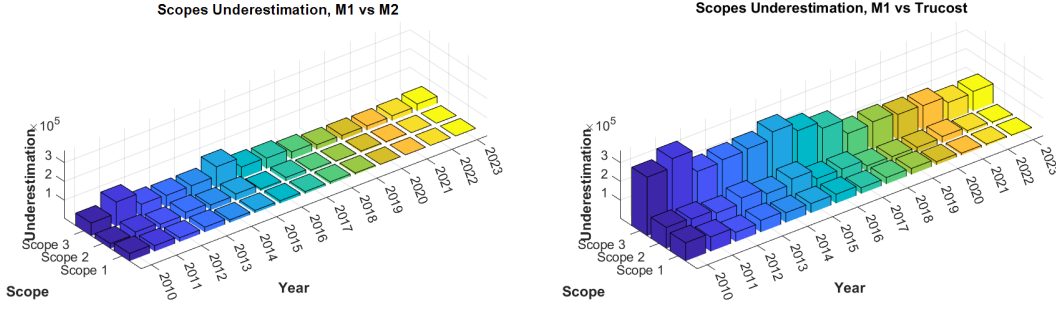(e) Trucost's prediction

Figure OA.5: Box plot of predictions

Figure OA.6: Underestimation in Scope Emissions

## OA.2   Kernel methods

We refer to Berlinet and Thomas-Agnan (2011) for introductory material related to reproducing kernel Hilbert spaces in probability and statistics.

### Reproducing Kernels

Define a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, for all $x_s, x_t \in \mathcal{X}$, satisfying

$$k(x_s, x_t) = \langle \phi(x_s) \,, \, \phi(x_t) \rangle, \tag{OA.2.1}$$

with feature mapping $\phi(x) = k(., x)$ that maps $x \in \mathcal{X}$ into some inner product space $\mathcal{H}$, called *feature space*. The feature space can be potentially infinite. Approximating a function $m(\mathbf{x})$ is challenging under this circumstance. Thanks to the representer theorem (Smola and Schölkopf (1998)), we can express $m(.)$ in terms of *kernel expansions* which takes the form,

$$m(x_t) = \sum_{s=1}^{T} \alpha_i k(x_s, x_t) = \alpha' k_t \tag{OA.2.2}$$

An important insight is that feature mapping $\phi(x) = k(., x)$ has the reproducing property, in the sense $\phi(x) = k(., x)$ spans the inner product space $\mathcal{H}$ which is called a *Reproducing Kernel Hilbert Space* (RKHS). Hence, any function $m(.) \in \mathcal{H}$ can be linearly spanned by $k(., x)$. In our context, $m(x_t)$ is a linear span of nonlinear transformation of the characteristic-based $\mathbf{x}$, namely $\phi(x_t) = k(., x_t) \equiv k_t$.

$\alpha = (\alpha_1, \cdots, \alpha_T)'$ is a weighting vector for a desired linear span.

There exists some nice properties in (OA.2.1) that are important to our context. First, since $k$ is symmetric, i.e., $k(x_s, x_t) = k(x_s, x_t)$, it can be seen as the metric for similarity in a nonlinear fashion. It encodes the similarity of high-dimensional variables at time point $s$ and $t$, and $x_s, x_t \in R^p$. Second, for $x_1, \cdots, x_p \in \mathcal{X}$, a kernel matrix, i.e., a real $T \times T$ symmetric matrix $K := \big(k(x_s, x_t)\big)_{s,t}$, is positive definite matrix, implying that $K$ is automatically invertible. Besides, the kernel matrix $K$ that encodes the similarity between any arbitrary high-dimensional variables at different time point is not limited to a linear structure, rather it encapsulates high-order moments similarity. The last nice property is the so-called "kernel trick" that links to (OA.2.1). If $x_s \in R^p$ lives in a high dimensional space, say $p = 500$, the corresponding feature map $\phi(x_s)$ that takes $p$ variables into a (infinite) feature space is spanned by their high-order moments. The kernel trick side-steps computational challenge by choosing a mapping $\phi(.)$ that leads to an easy-to-compute kernel function $k$. Instead of working on a tedious inner product of feature maps $\langle \phi(x_s) , \phi(x_t) \rangle$, it means that we can easily evaluate the corresponding kernel function $k(x_s, x_t)$.

## Some popular kernel functions

Within the kernel function class, the polynomial kernel and Gaussian kernel bring salient implications into asset pricing context. The polynomial kernel evaluates the similarity of a high-dimensional characteristic-based factor vector at two arbitrary time points in a feature space spanned by all monomials of degree $d$ in the input vector comprising of $p$ characteristic-based variables. For a polynomial kernel of degree $d$, it takes the form,

$$k(x_s, x_t) = \left(1 + \frac{x_s' x_t}{\sigma^2}\right)^d \tag{OA.2.3}$$

where $\sigma$ controls the contribution from higher-order terms, higher value of $\sigma$ less contribution from higher-order terms. (OA.2.3) corresponds to feature maps $\phi(x_s)$ that consists of all polynomials in the elements of a of degree at most $d$.

If there is no specific preference driven by prior knowledge of the true prediction function, the Gaussian kernel is a good candidate. The Gaussian kernel acts as a "catch-all" device as it never performs poorly than other ones (Exterkate (2013)), which explains why we choose it in our empirics. Because smart choices of feature maps $\phi(.)$ enable us to avoid exhaustive computations due to the curse of dimensionality, the Gaussian kernel can operate even if the feature space is infinite. Taking

inner product of the power series expansion of $\phi(x) = e^x$ leads to

$$k(x_s, x_t) = \exp(\frac{-1}{2\sigma^2} \|x_s - x_t\|^2), \qquad\qquad \text{(OA.2.4)}$$

where $\|.\|$ is the Euclidean norm. With respect to the frequency domain, the Gaussian kernel allows all frequencies (high and low) to be present (as opposed to polynomial kernels), albeit with very large penalties for high frequencies (considered as noise). The parameter $\sigma$ controls the roughness of the kernel. A higher value of $\sigma$ leads to a smoother kernel function.