

Master's Programme in Finance

Predicting ESG Controversies

Uncovering relationships between controversies and firm characteristics

Vernerri Suomela

Master's thesis
2024

Copyright ©2024 Vernerri Suomela

Author	Vernerri Suomela	
Title of thesis	Predicting ESG Controversies	
Programme	Master of Science in Economics and Business Administration	
Major	Finance	
Thesis advisor(s)	Prof. Markku Kaustia	
Date	Number of pages	Language
30.06.2024	60	English

Abstract

In this thesis, I examine the correlations between firm characteristics and future Environmental, Social, and Governance (ESG) controversies using Refinitiv ESG data. Specifically, I investigate how various levels of ESG scores, along with fundamental and stock market characteristics, predict future controversies one or two years forward. Employing logistic regression models, I regress dummy variables of future controversies on these explanatory variables.

The findings reveal significant correlations between ESG controversies and ESG scores, pillar scores, and most sub-pillar scores. Notably, higher ESG scores are associated with a higher likelihood of future controversies, potentially due to increased visibility or managerial opportunism, where companies inflate their ESG scores to gain the benefits associated with high ESG ratings. Additionally, firm size and past returns are significant predictors of controversies. The analysis indicates that English-speaking countries are overrepresented in ESG controversies, and industries with high public visibility or scrutiny, particularly those involving consumer products, experience more controversies.

To develop a more comprehensive predictive model, I employ an elastic net algorithm, for essentially a combination LASSO and Ridge logistic regressions for model selection and regularisation. This method yields more intuitive results, highlighting the relative significance of the different predictors like ESG scores, firm size, past returns, and country-specific factors in predicting future controversies.

Overall, this thesis contributes to the understanding of ESG controversies and their predictors, offering practical insights for investors and other stakeholders. The results suggest that while ESG scores alone may not effectively mitigate risks, a nuanced approach considering firm characteristics and industry context can enhance the prediction of ESG-related issues.

Keywords ESG, SRI, Controversy, prediction, corporate misconduct

Tekijä Vernerinen Suomela

Työn nimi ESG-kontroversioiden ennustaminen

Koulutusohjelma Kauppatieteiden maisteri

Pääaine Rahoitus

Työn ohjaaja(t) Prof. Markku Kaustia

Päivämäärä 30.06.2024

Sivumäärä 60

Kieli Englanti

Tiivistelmä

Tässä tutkielmassa tarkastelen yrityksen ominaisuuksien ja tulevien ympäristöön, yhteiskuntaan ja hallintotapaan (engl. environmental, societal, governance eli ESG) liittyvien skandaalien tai negatiivisten uutisten eli kontroversioiden välisiä korrelaatioita käyttäen Refinitivin ESG-dataa. Tarkemmin sanottuna tutkin, miten ESG-pisteytyksessä tilinpäätös- ja markkinadatan kanssa ennustavat tulevia kontroversioita yhden tai kahden vuoden päähän. Logististen regressiomallien avulla regressoin tulevia kiistoja kuvaavia dummy-muuttujia näiden selittävien muuttujien kanssa.

Tulokset paljastavat merkittäviä korrelaatioita ESG-kontroversioiden ja ESG-pisteytyksen välillä. Erityisesti korkeammat ESG-arviot ennustavat suurempaa tulevien kontroversioiden todennäköisyyttä, mikä saattaa johtua lisääntyneestä näkyvyydestä tai liikejohdon opportunistista, kun yritykset paisuttelevat ESG-pisteitään saadakseen korkeisiin ESG-luokituksiin liittyviä etuja. Lisäksi yrityksen koko ja aiemmat tuotot ennustavat merkittävästi kontroversioita. Analyysini osoittaa, että englanninkieliset maat ovat yliedustettuina ESG-kontroversioissa ja että toimialoilla, joilla julkinen näkyvyys on suurta, erityisesti kulutustuotteisiin liittyvillä toimialoilla, esiintyy enemmän kontroversioita.

Kattavampana ennustemallina käytän elastic net -algoritmia, joka on pohjimmiltaan yhdistelmä logistisia LASSO- ja Ridge-regressioita mallin valintaa ja regularisointia varten. Tämä menetelmä tuottaa intuitiivisempia tuloksia, joissa korostuu eri ennustetekijöiden, kuten ESG-pisteiden, yrityksen koon, aiempien tuottojen ja maakohtaisten tekijöiden, suhteellinen merkitys tulevien kontroversioiden ennustamisessa.

Kaiken kaikkiaan tämä tutkielma auttaa ymmärtämään ESG-kontroversioita ja niitä ennustavia tekijöitä sekä tarjoaa käytännön näkökulmia sijoittajille ja muille sidosryhmille. Tulokset viittaavat siihen, että vaikka ESG-pisteet eivät yksinään ennusta riskejä tehokkaasti, yrityksen ominaisuudet ja toimialan konteksti huomioon ottaessa monipuolisempi lähestymistapa voi parantaa ESG-ongelmien ennustamista.

Avainsanat ESG, SRI, kontroversio, ennustaminen, väärinkäytös

Table of contents

1	Introduction	6
2	Literature review	9
2.1	Effect of ESG controversies on stocks	9
2.2	Hypotheses formation.....	19
3	Data.....	22
3.1	ESG data.....	22
3.2	Explanatory variables	26
3.3	Sample.....	29
4	Methodology overview	33
5	Logistic regressions results	35
5.1	Methodology	35
5.2	Results.....	36
6	Elastic net regressions.....	46
6.1	Methodology	46
6.2	Results.....	47
7	Conclusion	52
	References.....	54
	Appendix.....	58

1 Introduction

The prevalence of using Environmental, Societal and Governance (ESG) factors in investing is growing rapidly with assets-under-management in ESG projected by Bloomberg Intelligence to reach 40 trillion USD by 2030 (Bloomberg, 2024). Still, the quality of ESG data available to investors is of questionable quality: ESG scores show high dispersion, as shown by Chatterji et al. (2016) and confirmed robustly in recent data by Berg et al. (2019), who show that correlations for ratings from 7 different rating providers were only between 0.38 to 0.71 in 2014 and 2017. Rating providers also change ESG scores retroactively (Berg et al. 2023), and Kaustia and Zhang (2023) show that higher ESG scoring companies in fact exhibit more ESG controversies.

Nevertheless, many investors are pouring capital into funds screened, weighted or portfolios otherwise formed based on ESG scores. Although the purpose for incorporating ESG aspects into investment decisions varies, if given that investors are looking to avoid ESG risks, one could quite reasonably expect that higher scoring companies carry less risk and would face fewer and less significant shocks. Kaustia and Zhang (2023) show the opposite to be true i.e. companies with high ESG scores face more shocks in the form of ESG controversies. The authors also consider scrutiny as a determining factor for controversies, showing that the scrutiny a specific industry faces is significant factor. The first part of this thesis follows a similar methodology with the same data source, and, as such, confirms their main findings. I go on to provide extensions to their work with a wider range of predictive variables and methods.

The issue of predicting ESG controversies has relevant implications for a large range of practitioners and academics. Controversies are known to have negative long-term (De Franco, 2020; Velazquez & Oliver, 2023) and short-term effects (Krüger, 2015; Capelle-Blanchard & Petit, 2019). Some studies have shown contradicting or nonlinear patterns regarding financial performance and controversies (Aouadi and Marsat, 2018; Dorfleitner et al., 2020). Based on the evident drop in share prices after an ESG controversy, being able to predict ex-ante which companies will have controversies and when would be a tremendous opportunity for portfolio and risk management as well as speculative trading. Abnormal returns can already be significant in long-term holdings (e.g. De Franco, 2020), and if the timing could be predicted, even greater compounded from short-term trading.

Moreover, by uncovering possible discrepancies in ESG scores, research like this could push rating and index providers to improve their methodologies. The differentiation of actual ESG risk and potential ESG impact from noise is important, both to investors looking to avoid negative shock and to the

environment and society at large. The results of this thesis speak for inconsistencies in how controversies are observed and reported as a part the Refinitiv ESG controversies score, which result in the score being of little use to stakeholders. A low-quality methodology also makes for a poor incentive for companies themselves to improve. Uncovering correlations between firm characteristics and future controversies could lead to more methodological theory-based research on individual interactions between controversies and firm, country or industry characteristics. Better quality ESG information like accurate risk metrics and predictors could lead to more accurate pricing regarding ESG risks. All the while, sustainability-valuing investors could improve the impact following selective allocation of their capital.

Controversies are not allocated randomly to companies: firm characteristics including ESG scores, firm fundamentals and stock market performance all have statistically significant implications for the number of future controversies. Some possible explanations for the results include varying visibility and scrutiny of companies, as well as mismatched incentives for management to seek out insurance-like effects related to active corporate responsibility. In this thesis, I demonstrate predictability in ESG controversies with relatively simple data mining methods, namely logistic regressions, and elastic net selection algorithms for regularized models. Specifically, the logistic regressions show positive correlations between Refinitiv ESG controversies and ESG scores as well as E, S and G pillar scores and their sub-pillars. Higher ESG metrics, especially the social pillar, predict a higher chance for future controversies. All logistic models control for firm size and have industry, year, and country level fixed effects. The sample is all public equities with Refinitiv ESG coverage from 2002 onwards, with 76,077 firm-year observations from 94 countries and 163 Global Industry Classification Standard or GICS sub-industries.

The elastic net models find metrics other than ESG scores to be stronger in standardised coefficients. Largest standardised coefficients are for revenue and total assets, both metrics of firm size. The direction of past returns is not as important as how much they stand out and attract attention to the firm: higher returns from two years ago predict more controversies as opposed to higher returns from past year predict less controversies. Different countries and industries also receive a significant loading in the model, with companies headquartered in Great Britain predicting the most controversies with the US as 5th worst out of 163.

To the best of my knowledge, the specific topic of predicting ESG controversies has not been studied in broader class of listed companies before, outlying a clear literature gap. The working paper by Kaustia and Zhang (2023) explores the topic and finds the same conclusions for ESG ratings as I do. As

said, I expand on their work with further explanatory variables on firm characteristics and with the elastic net model selection as a complementary alternative method more focused on prediction.

My contribution to this field is two-fold: I demonstrate that Refinitiv ESG scores are unreliable measures of a company's ESG riskiness. The observed controversies show a bias towards English-speaking countries, larger firms with exceptionally good or bad past performance, and visible sin-stock industries. Additionally, I provide evidence that ESG controversies are predictable using firm characteristics such as ESG scores, fundamentals, and stock market performance indicators.

The rest of the thesis is structured as follows: section 2 reviews the existing literature and section 3 describes the dataset. The general methodological approach is explained in section 4 whereas logistic regression models are described in section 4 and elastic net models in 5, with respective test set-ups and results included. Section 6 concludes.

2 Literature review

2.1 Effect of ESG controversies on stocks

As mentioned in the introduction, the effects of ESG controversies on stock prices and investor returns are still a matter of contention in financial literature. Moreover, the entire relation between corporate social performance or CSP and corporate financial performance is still very much undecided. Research with current data from recent years is hard to come by, and the results from past studies are inconclusive.

The setting is convoluted by constant change: the inflows into ESG investments are on a rapid rise with more funds, instruments, and different flavours of incorporating ESG criteria becoming available. Sustainability has become an all-encompassing aspect of society with its influence on the business world and financial markets clearly visible, especially in the EU with new legislation coming into effect on top of the clear demand for ESG assets from investors. Studies with datasets ending over five or ten years ago are therefore of questionable applicability to the financial markets of today, let alone future markets. The underlying mechanisms from past studies are still relevant, but, for example, one could quite reasonably expect the scrutiny companies face and the market reactions to ESG controversies to be rather different now as compared to ten years ago. The scrutiny and public pressure companies face regarding ESG issues is most certainly also higher now, which could impact the amount and magnitude of controversies.

Long-term effects of controversies for stock performance

For an understanding of the landscape around ESG controversies I first evaluate the performance of stocks after controversies.

The most recent results on the effects of controversies I found is from a commercial non-peer-reviewed article from Clarity AI (Velazquez and Oliver, 2023). Their dataset ranges from 2018 to 2022, which, given typical research-lag of a few years, is much more recent than peer-reviewed studies or even most working papers available online. Here the authors use artificial intelligence natural language processing models to identify 12,690 controversies and proceed to synthesize corresponding artificial counterfactuals for companies to evaluate the differences-in-differences in valuation. The average difference in market value after six months of the controversy treatment was -2.70% with severe controversies resulting in a drop of -5.39%. The Clarity AI article does not elaborate on whether the effect is an announcement shock immediately after the event, or a constant drift during the six-month period. A split based on the topic of the controversy shows that “Bad

governance” controversies caused the most significant average underperformance of -3.84% with environmental issues next at -3.35%.

The results from Clarity AI are consistent with previous academic research as well. A study by De Franco (2020) on data from 2010 to 2018 examines portfolios constructed quarterly based on a controversy metric derived from Sustainalytics ESG controversy ratings. The portfolios are formed using ratings from available one day before the quarterly rebalancing therefore only using ex-ante information and capturing post-announcement drifts and not the announcement effects around the news breaking. The author shows that high controversy portfolio of large to mid-capitalisation European stocks had annual returns of 1.85% p.a. during the sample as opposed to the 7.14% p.a. returns of the unfiltered benchmark index. The results are similar for US stocks but interestingly the results do not hold for Asia-Pacific markets, where the high controversy portfolio in fact out-performed the index 8.05% to 5.77% p.a. respectively. However, the high controversy portfolio was not broadly diversified consisting of only ten stocks. The differences between geographies are still very striking, hinting at a possibility for a vast difference in how controversies are priced or what constitutes a controversy.

Aouadi and Marsat (2018) find similar, contradictory results in a sample from 2002 to 2011: ESG controversies are correlated with higher valuations measured by Tobin’s Q, even when controlling for a vast amount of firm level factors including CSP scores (Corporate Social Performance scores, the predecessor to ESG scores), size, profitability, capital structure and so on. The author regress logarithmically scaled Tobin’s Qs with a dummy indicating whether a company is involved in an ongoing controversy based on Asset4-Thomson Reuters (currently branded as Refinitiv) controversy scores and CSP scores. The loading of the controversy scores turns insignificant when an interaction variable of CSP scores with controversy scores is added. The interaction receives a coefficient both high in magnitude and significance, indicating that companies with high CSP scores yet involved in controversies have higher valuations. This could be a reverse causality: higher CSP companies with high valuations attract more attention and have more analyst coverage leading to more controversies.

A study on similar Refinitiv data but a longer period ranging from 2002 to 2018 by Dorfleitner et al. (2020) shows that portfolios based on a 10% best and worst controversy scores both had significant excess returns in a long-term Fama-French five factor model (Fama and French, 2015). The correlation between controversies and performance seems to be U-shaped with both highest and lowest quantiles outperforming the market at large.

Announcement shocks and other short-term effects

Event-studies around controversies show drops in valuation with commonly cited effects of -1.31% cumulative abnormal returns in a window of [-10,10] i.e. 21 days around the event reported by Krüger (2015) and -0.1% cumulative abnormal returns (CARs) around negative ESG news in a shorter [-1,1] window reported by Capelle-Blanchard and Petit (2019). Differences in the definitions and sources for controversies may account for the large difference in magnitude with Capelle-Blanchard and Petit evaluating all negative news sourced from Covalence EthicalQuote from 2002 to 2010, a much broader pool, and Krüger using KLD from 2001 to 2007 as his source.

A more recent figure from the US is reported in a working paper by Cui and Docherty (2020) ranging from 2000 to 2018. For a sample of S&P 1500 Composite constituents, they report a highly significant -0.773% Carhart (1997) four-factor CAR around a [-10,10] window. The subgroup of smaller stocks (S&P SmallCap 600 constituents) exhibited a much more dramatic shock with a drop of around 2%. The size effect was rather consistent with S&P MidCap 400 constituents dropping close to 1% and S&P 500 constituents dropping by about 0.5%. Positive news did not result in significant CARs for any groupings.

Intuitively, due to the definition of a controversy as negative news, the announcement effect is negative. For speculative trading and risk management, predictive tools should focus on the more significant controversies with CARs of over -1%. For bigger stocks, the value-destroying effect is of course large in absolute size, and, for value weighted portfolios, the value-destroying effect can be economically significant even considering smaller drops. The literature covering the ex-post characteristics of financially significant controversies is much broader as almost all the studies discussed in this section shed some light into the matter.

Determinants of impact magnitude as cues for future controversies

Although providing backing the importance avoiding ESG controversies, the stock market reactions explored above are only of secondary importance to this thesis. My main research question is what causes ESG controversies and what kind of companies face extraordinary amounts of controversies. Nevertheless, the prior literature on stock market reactions and financial performance does shed light on what makes controversies material and significant. The significance and magnitude of the announcement effects could be a proxy for how visible stocks are and how much scrutiny they face from investors, analysts, the media, non-governmental organisations or any other related parties. This would in turn be a driver for controversies as well. The different hypotheses and tests of previous studies on markets reactions with different controls and sample splits are more interesting to the topic of

predicting ESG controversies, than results broadly evaluating performance during ESG controversies as a whole.

Aouadi & Marsat (2018) show that ESG controversies themselves are in fact correlated with higher stock valuations. They explore different hypotheses on the underlying mechanisms and interactions with firm size and visibility. The authors present results after splitting the sample in two based on various metrics. Splitting was done on stock fundamentals related to visibility, namely ROA and size, and as well as four other metrics: splits based on the median of press freedom index in the country of headquarters, Google search volumes, analyst coverage and a binary division on whether the company has received a corporate social responsibility (CSR) award. After robustness checks and addressing endogeneity with lagged variables and a two-stage least squares model, the authors conclude that a visibility and the amount of scrutiny a company faces could explain differences in the magnitude of the reaction to controversies. Larger, well performing companies with more analyst and media attention located in countries with greater freedom of press are more likely to have controversies, which lower their controversy scores despite their high CSP scores.

ESG scores and the controversies are, at least in the Refinitiv data, strongly correlated with firm size: bigger firms have higher ESG scores (e.g. Dremptic 2019), but more controversies (Kaustia and Zhang, 2023). The materiality of controversies that larger firms face is questionable based on the event study results.

Krüger (2015) reports on event-window CARs with a similar approach of various groupings for the types of news and controls for companies. Of negative news i.e. controversies, all news average at -1.31% in a [-10,10] window, but the biggest effects are for community flagged news with -3.33%, followed by environment and product with -3.03 and -1.22% respectively. In regressions of CARs, a control variable for S&P credit ratings is significant with a negative effect i.e. companies with a higher credit rating suffer more from controversies.

For stock market reactions and lasting effects, size does not appear to be a clear determining factor. Krüger (2015) reports an insignificant loading for logarithmically transformed market capitalisation and Cui and Docherty (2020) show that smaller capitalisation companies experience smaller negative CARs. However, Capelle-Blanchard and Petit (2019) show a negative loading for log total assets, meaning bigger companies have bigger draw-downs after controversies. In Aouadi and Marsat (2018), the main finding of the ESG controversy score in interaction with CSP scores being significant to valuations only holds for the larger half of stocks after a split based on size.

In contrast, after splitting their sample by market capitalisation, Dorfleitner et al. (2020) report stronger effects for the smaller sample and for equally weighted portfolios. The split test is, however, insignificant for the portfolio of worst scoring 10% in terms of ESG controversies. The alphas remain positive and significant for the highest quantile portfolio, and the effect is biggest in the equal weighted portfolio of the below median market capitalisation group.

The effect of size can be somewhat misleading, if one can assume that bigger companies have more news coverage and therefore more negative news. In that case, a single piece of news is less likely to affect stock prices, and therefore event-studies focusing on CARs around the announcement would have smaller CARs due to dilution. Studies considering ESG controversy scores would, however, pick up on the overall volume of controversies. Although the increase in controversies with size is evident, differences in the sources and definitions of negative news and ESG controversies vary further convoluting the assumptions that can be drawn regarding the significance of controversies for bigger companies.

Current studies on ESG scores and controversies

Outside of the effects on stock prices or other investor returns, the topic of the causes of ESG controversies has not been very widely studied yet.

A working paper on ESG scores and their relation to ESG controversies by Kaustia and Zhang (2023) finds a negative correlation between Refinitiv ESG scores and pillar scores and controversies. Their tests show that higher ESG scores and individual E, S and G pillar scores with a two-year lag all correlate robustly with more future controversies. A control variable of log market capitalisation is also consistently highly significant and large in magnitude. Their sample consists of 21 countries from Europe and the US, after dropping smallest countries from the sample. The authors also test Refinitiv controversies against MSCI ESG scores and show that the same does not hold. Higher MSCI ESG scores as well as S and G scores separately correlate with less controversies. The environmental pillar is inverse as with Refinitiv ESG scores. The effects of salience and scrutiny are also tested and show that companies with more salience i.e. more disclosure face more scrutiny i.e. media attention. When media scrutiny is high, more controversies are reported for companies with high ESG ratings.

A recent study by Agnese et al. (2023) explores a similar topic as I do. The authors examined the correlation between ESG controversies and governance scores in the banking sector. Their study found in a sample of some 567 listed US and EU banks that higher Refinitiv ESG governance pillar scores and the underlying sub-pillar scores correlated positively with the ESG

controversies scores, meaning good governance correlated with less controversies. However, similar studies on non-financial companies have not been published. When sampling all financial sector listed companies including investment trusts and not just banks, I find controversies as reported by Refinitiv to be notably less common compared to the just the banking sector. The results may not be extendable to the whole financial sector let alone other, nonfinancial industries, as other real economy sectors function rather differently. For financial companies, the impact they have on the environment or on society is unlike that of others, influencing ESG mainly with financing and capital allocation, and less their own operations.

Studies on misconduct and good governance

Previous research has established some connections between firm characteristics relating to governance and the likelihood of encountering ESG controversies or corporate misconduct. Liu (2016) identifies a link between a corporate culture of corruption, a measure to capture general attitude towards opportunism, and the prevalence of misconduct. Similarly, the duration of a CEO's tenure is found to influence corporate behaviour, with Altunbaş et al. (2018) noting that longer tenures can contribute to unethical practices in banks. Additionally, Chen et al. (2018) highlight that the foreign residency of the controlling person within a Chinese firm is associated with increased incidents of corporate misconduct.

Conversely, Neville et al. (2019) assert that board independence plays a crucial role in reducing unethical behaviour within firms, and Liu (2018) reveals that greater female representation on boards and the presence of female CEOs are associated with a lower incidence of environmental violations. This suggests that gender diversity within leadership can contribute to better ESG outcomes and reduce the likelihood of corporate misconduct. Gelman et al. (2021) suggest that financial advisory firms with higher market power are less likely to engage in unethical practices.

All in all, previous literature seems to show that good governance is related to less controversies. Metrics like board cultural and gender diversity, and the independence of the board as well as other nominated committees are used in the Refinitiv ESG methodology to determine the governance and ESG scores. Therefore, the higher ESG scores or governance pillar scores should reasonably predict fewer controversies.

Factors determining ESG scores

An analysis by Drempetic et al. (2020) shows that Refinitiv ESG scores are correlated with size, the amount of resources a company uses for ESG data, and the amount of data a company has either disclosed themselves or third-parties have revealed.

It is commonly noted that the ESG pillars and sub-pillars are highly correlated, as is confirmed in tables 1 and 2 of section 3 of this thesis. This hints at companies being somewhat dichotomous: some being responsible or green, and others being indifferent or irresponsible so-called brown companies. This may arise due to some companies being less active in voluntary ESG disclosure with less resources allocated to ESG, a joint determination reflecting the effect of size and the amount of third-party attention. That is, if other companies are bigger and attract more attention from third parties, they are more active in voluntary disclosure and allocate more resources to generate and disclose ESG data, they will subsequently have higher ESG scores. On the other hand, and especially historically with fewer ESG-conscious investors, suppliers, customers or other stakeholders, companies may have been inclined to accept their below-average position as compared to peers, and to not cater to ESG-conscious stakeholders. If a company does not see a competitive advantage or other return for their investment into ESG, the dichotomy into green and brown is logical, resulting in high correlations across different ESG metrics.

Investor biases in evaluating the effects of controversies

Overall, the materiality of ESG information is often hard to gauge due their soft and qualitative nature. Quite often, the negative news cannot be assigned a defined amount a firm's value is expected to drop due to revenue loss, increased costs, fines etc. The information is slowly incorporated into the stock price, taking multiple days for the announcement to be priced in and even still stock prices often tend to decline for months afterwards. As controversies leave a lot up for interpretation and subjective assessment, they also present a potential for biases to arise.

Capelle-Blanchard and Petit (2019) test for both physical distance as well as shared language as determinants for the CARs during controversies. The shared language dummy, defined as 1 if the controversial event takes place in a country with the same language as the headquarters of the firm and 0 otherwise, receives a significant negative loading in all $[-1,1]$ window regressions the authors present. Investors themselves may be subject to attention bias and allocate more sentimental weight to these events, or the correlation may simply arise from increased visibility i.e. media and analyst coverage, since primary sources are available in the same language. This would fit in with the attention narrative as a driver of significance and the overall number of reported controversies.

Another commonly noted pattern is that positive news are not priced in as clearly as negative news (e.g. Capelle-Blanchard, 2019; Krüger 2015). Again, this could arise from investors themselves having loss aversion or that

negative news make for more salient headlines, given the loss aversion of consumers. A commonly cited ballpark rule-of-thumb is that negative news can be twice as powerful as positive news (Tversky and Kahneman, 1992).

Cui and Docherty (2020) examine return after ESG controversies from a salience theory perspective. They show that investors overreact to negative ESG news, and that returns typically mean-revert over a period of 90 days. The sample is on S&P Composite 1500 constituents from 2000 to 2018. The authors do note that the evidence is driven by small stocks, where the limits to arbitrage are higher and mispricing is more prevalent. S&P SmallCap 600 constituents reverted from an initial drop of over -2% CAR around a day or two after the controversy to around -0.6% at 90 days after. The results are interesting when compared to the returns reported by De Franco (2020), who shows spread exists in quarterly rebalanced portfolios formed according to past ESG controversies in a sample of large and mid-capitalisation stocks. For the constituents of the Solactive GBS US index, De Franco reports a spread from 14.50 % p.a. returns from a no controversy portfolio to 7.42% in a high controversy portfolio, with all other portfolios (low, moderate and unfiltered benchmark) consistently in between.

Biases may influence the perception of controversies and subsequently effect the number of cases that are noticed, deemed to surpass the threshold of significance required to be classified as a controversy and reported. The salience could lead overreactions especially in industries that are highly visible in society and media regarding their sustainability, like sin-stock, fossil fuel or more consumer-facing industries like automotive manufacturers or airlines.

From an investor's standpoint, the implications of controversies can be coarsely divided in two effects: reputational market reaction shock and the shocks future cash-flow generating potential. These are of course overlapping, but when considering potential ESG controversies which go unnoticed, the reputational issues do not arise, yet the source of the controversy may have other financial effects or simply delayed effects. For example, issues with local workforce may not make their way to (international) news but they may have significant effects on productivity and costs for the company. In countries with limited freedom of press, international media or analyst coverage, more controversies may go unnoticed, or they may be suppressed. In evaluating the ESG controversy risk in such countries, the predicted number of controversies is not directly comparable to countries without such issues, like companies headquartered in the US and the EU. That is to say that the number of reported controversies recognised by rating agencies like Refinitiv is not equal for all companies in different geographies, industries or other groupings which affect their visibility in global media and analyst coverage. The controversy scores do not capture the whole extent of the issue, and it

leaves room for interpretation by stakeholders in determining the impact of a reported controversy given firm specific characteristics.

Insurance-like effect of CSR activities before controversies

A study on the actions and performance of a company following a controversy by Li et al. (2019) shows a rise in symbolic, non-substantive CSR activities following ESG related shocks. Moreover, the symbolic CSR is shown to improve the firms' reputation and the stocks' performance.

Already before the widespread adoption of ESG, studies like Godfrey (2005) and Peloza (2006) show that CSR activities can have an insurance-like effect. Companies at risk are more likely engage in CSR to build social credibility and so-called moral capital to mitigate future controversies.

Ferrés and Marcet (2021) show that companies under investigation for price fixing more commonly engage extra CSR when the investigation is launched. The positive CSR is in fact related to reduced fines and a smaller decline in sales after the cartel is exposed.

More recently, Shiu and Yang (2017) reassess the insurance-like effect broadening the scope to bonds as well as stocks. In line with previous results, they show that stock and bond prices (although bond price reactions are not statistically significant) drop less for companies which had long-term CSR engagements prior to the controversy. In accord with Godfrey (2005) the authors conclude that CSR must be perceived as genuine and long term to avoid the perception of opportunism or ingratiating. The insurance effect also holds only for the first controversy a company faces, whereas for following controversies the effect is much smaller and not statistically significant.

Capelle-Blanchard and Petit (2019) use a greenwashing component in the regressions of CARs surrounding negative ESG news. They define it as the percentage of positive ESG news published by the company themselves out of all of the positive ESG news related to that company i.e. published by the media, NGOs or other third parties. The correlation is significant and positive with companies which had reported more about their own positive ESG actions suffering a 1.8% smaller hit to their stock prices around a controversy.

Given the insurance-like effect, company management would be incentivised to engage CSR if they themselves saw risks for controversies. Management could have inside information on the misconduct, negligence, improper risk management or other ESG related issues. They would then strive for CSR and high ESG scores to mitigate the effects of controversies. This could explain a portion of the findings of Kaustia and Zhang (2023), who demonstrated that high ESG companies faced more controversies. Especially the difference

between MSCI and Refinitiv ESG scores is meaningful: Kaustia and Zhang (2023) note that Refinitiv scores are more based on disclosure by the company themselves but, MSCI considers outside sources like information made public about the firm by others. Refinitiv score would therefore be more open to selective disclosure or even manipulation as firms know the metrics used in the evaluation.

Trust and penalties for non-compliance

To evaluate the theoretical background of compliance and reporting, literature more focused on financial fraud and non-compliance serve as a good parallel for ESG reporting and controversies.

Karpoff (2021) evaluates the financial fraud from the perspective of theoretical constructs centred around trust and enforcement to compliance. He considers two models: the Trust Triangle formulated by Dupont and Karpoff (2020) and a contractual enforcement model by Klein and Leffler (1981). Both models consider how opportunistic sellers could sell bad quality stock or financial claims for financial gain. The Trust Triangle model considers three types of enforcement: Third-party enforcement like laws, regulations and institutions; Related-party enforcement like market-forces and reputational drivers and lastly First-party enforcement for personal ethics, integrity and culture. The Klein and Leffler model formalizes the market forces and reputational capital as a mechanism to penalize fraudulent behaviour. All in all, the reputational risk of being outed as fraudulent or otherwise bad outweighs the benefits of fraud, even if the third-party regulatory bodies do not punish the misconduct.

From an ESG standpoint, controversies are mechanism for outing irresponsible firms from posing as responsible. Managers have ample incentives, both personal and as agents for stockholders, to inflate ESG metrics if possible. If done within the law but still with a condemnable amount of dishonesty, the only risk is being exposed by the related-party enforcers (media, analysts, rating agencies, NGOs or other stakeholders), or possibly other, responsible managers or shareholders due to their personal first-party enforcement. Such news would constitute ESG controversies for the company, providing a possible connection between high ESG scores and high levels of controversy. On the other hand, companies providing transparency by reporting honestly and openly as well as voluntary disclosure may have lower ESG scores consequently, but coincidingly less ESG controversies.

2.2 Hypotheses formation

The main goal of this thesis is to evaluate, what characteristics companies facing extraordinary amounts of controversies share. I start with the assumption that ESG scores should measure ESG risks accurately. To formalise the approach, I consider the following null hypotheses:

H0: Companies with better ESG performance have less controversies

As discussed in the introduction, investors would typically assume that high ESG scoring companies would have fewer negative shocks arising from ESG issues. However, Kaustia and Zhang (2023) show the opposite to be true. The true mechanism between ESG scores and controversies is difficult to uncover, but current evidence points toward a clear negative correlation even when controlled for size, geography, and industry. Companies, which are looking to give a better image of themselves to stakeholders and be more active in ESG disclosure, will have higher ESG scores. This could spur more attention and scrutiny, resulting in more controversies uncovered.

Given the insurance-like effect, managers with insider information about the firm's activities could engage in CSR activities to mitigate the effects of controversies. This could lead to ESG controversy risky companies to inflate their ESG scores.

Similar to financial fraud, dishonest companies may try to sell false claims to investors and rating agencies to improve their ESG scores, which in itself is a governance issue. As governance issues are often correlated with other ESG issues, these firms may also be more likely to behave irresponsibly elsewhere leading to controversies. Controversies are, in a way, the mechanism to expose companies posing as more sustainable and responsible than they actually are: dishonest companies will push their ESG ratings until a controversy is discovered leading to a correlation between high ESG ratings and controversies.

To further evaluate ESG performance or more precisely overall environmental sustainability with clearer, more objective metrics, I use carbon equivalent emissions amounts relative to revenue to gauge the carbon efficiency of companies. This measure should be less subject to the opinion of the rating agency or data vendor, and firms should have less opportunities to manipulate the reported emissions. Scope 3 indirect emission are, however, very difficult to calculate with little widespread adoption of established standard practices.

I also evaluate the effects of the different ESG pillars and sub-pillars. Krüger (2015) shows that the magnitude of controversies on valuations varies by the type of controversy. The categories with the largest drops significant at least 5% in a [-10,10] window are Community with -3.33%, Environment -3.03% and Product with -1.22%. With potential behavioural effects like attention bias and irrationality in evaluating what constitutes a controversy, the limited attention could be overweighted in companies claiming to be responsible in these sectors.

Considering the potential effect arising from managerial dishonesty as described above, some ESG sub-pillars may be easier to affect and improve compared to industry averages. Therefore, these sub-pillar scores will be higher when a controversy is revealed.

Visibility as a convoluting factor

ESG controversies should ideally be about objectively identified issues and material shocks. However, controversies are not reported equally across firms but aspects like visibility in society and media, analyst coverage and investor attention will affect what companies are reported as having ESG controversies.

To start with perhaps the most salient or rather obvious hypothesis, larger firms as measured with any relevant metric like total assets, market capitalisation or revenue will have a higher number of controversies. The larger the operations of a company, the more chances for controversies to arise. But compensating linearly for the size of a company is not enough, that is, the rate of controversies does not double as the size (e.g. revenue or assets) of the company doubles. Larger companies face much more scrutiny over their actions, attracting attention not only from analysts or reporters but also from activists and NGOs with a specific agenda of improving the ESG impact companies have and uncovering potential misconduct. Similarly, stock with extraordinarily high returns within recent years would receive more media and analyst coverage. Bigger companies may also be more active in voluntary disclosure as a part of their ESG activities, which could explain improved ESG scores but also exacerbate their number of controversies.

In line with the attention theory, companies in industries with high levels of societal attention and scrutiny should be more likely to have controversies. On top of typical sin-stock industries like tobacco or gambling, other high attention industries, especially in terms of ESG, could include fossil fuels, automotive manufacturers, aviation, and consumer goods. On the other hand, business-to-business industries like intermediate products or components without consumer products would face fewer controversies.

Previous studies (e.g. De Franco 2020), have shown that, for example, companies head-quartered in Asian countries have very different stock price reactions when involved with controversies as compared to US or European companies. Refinitiv markets their ESG controversy scores as based on “Global media” (LSEG 2022), yet some studies using the precursor ASSET4 database describe it as collected from English-speaking news (e.g. DasGupta 2022). Despite the global breadth of news evaluated in determining controversies, US and Western European companies are more visible in international media and arguably held to a higher standard as forerunners in ESG, especially EU countries. Due to biases, news in a shared language may be more salient and thus more commonly brought to light and classified as controversies.

The overallocation of controversies for companies reporting primarily in English may also be due to the process of identifying controversies, which, although not disclosed by Refinitiv, is most likely done using algorithmic or AI language processing to scan news publications. These models may be better trained for English as compared to other languages, therefore potentially adding to the overweighting. This effect could be extended to a lesser extent to other globally common languages like Spanish.

The freedom of press and the thresholds for what is deemed worth reporting and what is classified as a controversy may vary greatly. For example, China is currently second to last in the Press Freedom Index (Reporters Without Borders, 2023) losing only to North-Korea. Therefore, despite the large economy, high stock market capitalisation, questionable development in ESG matters compared to the EU or the US, still I hypothesise less controversies for firms headquartered in China. On the other hand, Western European and Northern European countries are commonly on the top of the leaderboard for the Press Freedom Index.

Overall, the logistic regression tests aim to first reject the null hypothesis and the predictive test in this thesis aim to reconcile which factors are the most important in determining future controversies.

3 Data

I use the entire sample of available data from Refinitiv, mostly restricted by their ESG coverage. For this section and the tables within, I describe the largest sample used in model 1 of table 8, a regression of a one-year forward controversy dummy on ESG and fundamental right-hand variables. This sample starts from 2002, that is, first predictive variables are from 2002 and first controversy scores from 2003. The sample was collected during February 2024, when most of 2023 ESG ratings were not complete. Therefore, the sample is limited to predictive variables mostly from 2022 and outcome controversy dummies from 2023. For two-year forward models, the logic is the same, but the most of the latest firm-year predictors are from 2021 and outcomes from 2023, resulting in a smaller sample size.

3.1 ESG data

Refinitiv ESG scores

For ESG data, I use data from LSEG Refinitiv, formerly also known as Thomson Reuters or ASSET4. Their ESG data is commonly described as transparent with an at least relatively clear methodology and good coverage (eg., Dorfleitner et al. 2020, Agnese et al. 2023).

Coverage in terms of firm-year observations for the sample period is presented in Figure 1. The trend is clearly rising, as one would expect given the ongoing rapid growth in ESG investing. The distribution of firm-year observations is therefore organically weighted more on recent years. The ESG evaluations for 2023 were mostly missing as of data collection in February 2024.

Refinitiv ESG scores are comprised of three pillar scores in each Environmental, Social and Governance respectively. These are further divided into 10 sub-pillars, which are in turn calculated based on a maximum of 186 data-points mostly from voluntary disclosure by the company themselves. All the levels from sub-pillars to ESG scores are evaluated on a percentual scale from 0 to 100 with a theoretical median of 50. The percentages are grouped and calculated on an industry level resulting in essentially best-in-class industry-neutral, comparable ratings.

Controversy scores

The ESG controversies score is also ranked based on quantiles starting from 100 equating to no recent controversies to 0. The scale from <100 to 0 is a percentual score within the industry of the company. For example, a score of 50 shows that the company is at the median of all companies with controversies and thus an imperfect score i.e., a score under 100. The ranking is based

Figure 1: Time series of the number of public stocks with an available Refinitiv ESG score for the given year.

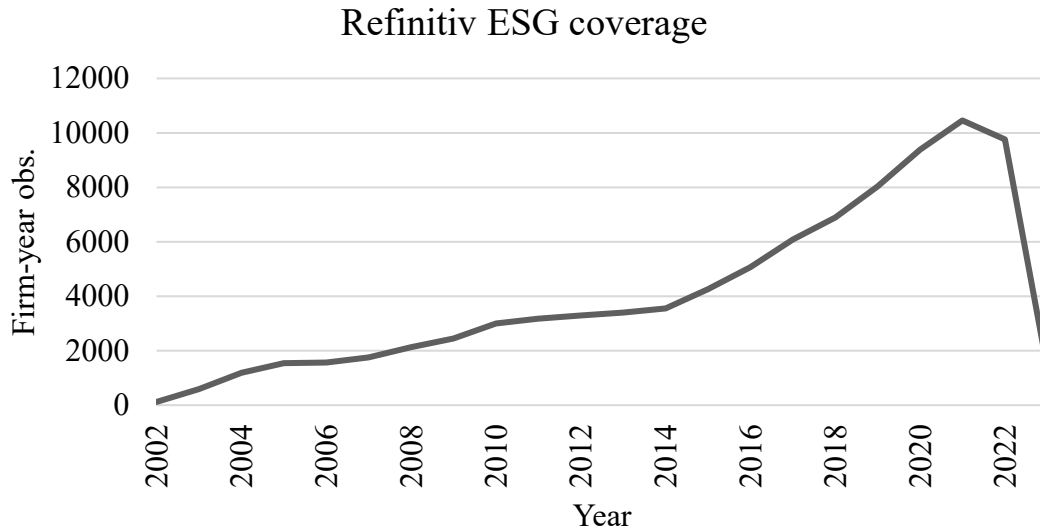


Table 1: Summary of ESG controversies scores.

	Observations	Mean score	Standard deviation	Min	Max
Whole sample	76,077	92.06	21.24	0.42	100
<100	13,004	51.09	27.87	0.42	99.15

Table 2: ESG pillar score correlation matrix

	Environmental pillar	Social pillar	Governance pillar	Log total assets
Environmental pillar	1.00			
Social pillar	0.72	1.00		
Governance pillar	0.41	0.42	1.00	
Log total assets	0.50	0.37	0.29	1.00

on controversies the firm has experienced recently with the duration of the controversy relating to how much aftermath such as lawsuits or ongoing media coverage the issue receives. Overall, the controversies score is well suited for academic studies, as it is marketed as industry neutral and based percentage rankings (LSEG, 2022), which makes comparing companies more insightful.

Table 3: ESG sub-pillar correlation matrix.

	Resource use	Emissions	Environmental innovation	Workforce	Human rights	Community	Product responsibility	Management	Shareholders	CSR strategy
Resource use	1									
Emissions	0.82	1								
Environmental innovation	0.50	0.49	1							
Workforce	0.71	0.70	0.37	1						
Human rights	0.65	0.59	0.39	0.54	1					
Community	0.45	0.41	0.25	0.46	0.43	1				
Product responsibility	0.51	0.49	0.37	0.48	0.43	0.34	1			
Management	0.27	0.27	0.16	0.28	0.23	0.27	0.19	1		
Shareholders	0.12	0.11	0.07	0.11	0.09	0.12	0.07	0.19	1	
CSR strategy	0.73	0.73	0.45	0.64	0.56	0.40	0.44	0.29	0.12	1

Given the rather soft, qualitative nature of text analysis, the threshold for what constitutes a controversy can be quite imprecise. The Refinitiv documentation (LSEG 2022) provides little information on what criteria is used to identify controversies, outside of categorisation principles. The documentation does address size bias, stating that large-cap companies are compensated for the additional media attention they attract compared to smaller companies. Specifically, companies with market capitalisations of over 10 billion USD are classified as large, receiving a weighting of 0.33 for controversies they face. For companies with capitalisations of under 2 billion USD, the weight is 1, and for mid-cap companies falling in between, the weight is 0.67.

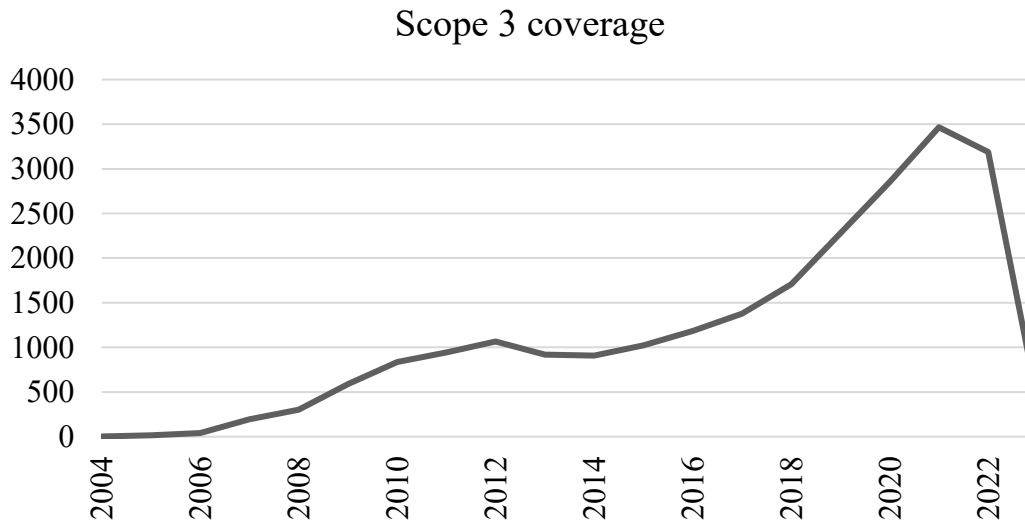
For many of the tables, I include a “severe controversy” model. Here, the outcome variable is a binary dummy for the firm-year observations, where the controversy score is below 50. This indicates a stronger controversy as compared to the standard outcome of any observed controversy.

Emissions data

To evaluate environmental effects of companies outside of the Refinitiv ESG metrics, I use CO2 equivalent greenhouse gas emissions using all three scopes. Scope 1 is emissions directly emitted by the company or subsidiaries that the company has control over. Scope 2 is from purchasing energy, namely electricity, heating, and cooling. Scope 3 is from indirect sources from both up and downstream from the company, mainly consisting of suppliers and customers of a company.

Especially scope 3 emissions are hard to evaluate, and no clear best-practice or standard methodology has been established leading to missing data and bad quality data with large variance source by source. The quality and coverage have increased recently and some initiatives like the Carbon Disclosure Project exist to unify the methodology. I selectively use the CO2 equivalent data in models, which in turn limits the number of observations available. Figure 2 plots out the coverage of scope 3 data availability.

Figure 2: Time series of the number of public stocks with coverage of both Refinitiv ESG metrics and scope 3 CO2 equivalent emission data for the given year. Fall-off for 2023 is due to unfinished and missing data as of collection in February 2024.



3.2 Explanatory variables

To explore the different hypotheses and uncover correlations between firm characteristics and future controversies, I use a wide range of models with different variables of interest. The selection is based on the literature review including firm characteristics, which have been shown to have some effect on ESG controversies, misconduct or CSR. Some key fundamentals are included to proxy for size, visibility, stock market performance and profitability.

Metrics to proxy for portfolio management risk factors, namely the Fama & French factors, are also included in the predictive models. Asset pricing models like the Fama & French factor model claim to capture relevant risk characteristics arising from distress, that would otherwise be unobserved in a simple one-factor capital asset pricing model. This unobservable distress captured by, for example, the high-minus-low long-short portfolio could be correlated with increased risk regarding ESG controversies as well. To mimic these known asset pricing anomalies in the predictive models, I use a variety of fundamental and market variables. The related risk factor is in parentheses in table 5. Small-minus-big (SMB), high-minus-low (HML), conservative-minus-aggressive (CMA) and robust-minus-weak (RMW) refer to the Fama-French (Fama and French, 2015) five factor model.

As the ESG scores are only updated yearly, capturing the Carhart (1997) momentum effect cannot be done as effectively. The timing of the ESG rating is distributed seemingly randomly throughout the year whereas the fundamentals for a given financial year are updated at the turn of the year. Refinitiv ESG scores and controversy scores are also at times revised and restated (Berg et al 2023). The only opportunity was to use the total return of the previous year and the year before that. For example, a company receiving a ESG rating in June 2023 for the financial year 2022, that firm-year observation row will have a total return datapoint for the financial year 2022 and a lagged return for 2021.

The Refinitiv Workspace has good coverage on fundamentals and market data, with the most common control of market capitalisation only leading to the omission of 190 observations from the 89,040 firm-year observations available with ESG coverage starting from 2002. The number of observations is reported for each of the models.

For assessing hypotheses 2 and 3, I use Refinitiv ESG scores including their subscores, subscore pillars and other more granular metrics like carbon intensity.

Variables that are used without a logarithmic transformation are winsorized to the 1st and 99th percentiles. This is to extreme eliminate outliers, which without winsorisation had significant effects on some correlations. Coincidentally, lots of firm-year observations that are winsorized would have been eliminated in regression models anyway due them to missing data for an included variable, so the amount of winsorized observations is less than 2% for all models. Winsorized variables include PE, EBITDA, past returns, revenue, CAPEX, common equity, total assets, and market capitalisation. For these, absolute values used in models are winsorized and ratios such as such as leverage, book-to-market, CAPEX-to-revenue are calculated from winsorized variables. Logarithmic variables are not calculated from winsorized variables, as outliers have an insignificant effect after the transformation and the winsorized cluster at the 1st and 99th percentiles would in turn effect results in an undesired way.

For industry dummies, I use GICS subindustry level codes with 163 different categories. Given the size of the dataset, only a handful of industries are disregarded in the logistic models due to perfect predictors, that is, the industries do not have any firm-year observations with controversies.

Table 4: Correlation matrix for fundamental and market variables. Time t is the year when the prediction is made.

	Contr _{t+2}	Contr _{score}	ESG _{score}	Ret _{t-1}	Ret _{t-2}	Mk.Cap	Log revenue	Log capex	Capex/revenue	Book/market	Leverage	PE	EBITDA/revenue
Contr _{t+2}	1												
Contr _{score}	-0.326***	1											
ESG _{score}	0.219***	-0.258***	1										
Ret _{t-1}	0.025***	-0.009***	0.007**	1									
Ret _{t-2}	0.016***	0.028***	0.000	-0.002	1								
Log Mk.Cap	0.308***	-0.272***	0.431***	0.186***	0.166***	1							
Log revenue	0.314***	-0.306***	0.460***	0.022***	0.037***	0.723***	1						
Log capex	0.294***	-0.279***	0.419***	0.002***	0.020***	0.681***	0.771***	1					
Capex/revenue	-0.003	0.003	-0.012***	0.010***	0.003	-0.019***	-0.063***	-0.009***	1				
Book/market	-0.005	-0.015***	-0.009***	-0.062***	-0.052***	-0.126***	0.010***	0.021***	-0.001	1			
Leverage	-0.098***	0.106***	-0.134***	-0.005	0.037***	-0.068***	-0.235***	-0.142***	0.002	0.015***	1		
PE	-0.014***	0.017***	-0.047***	-0.019***	0.035***	-0.021***	-0.122***	-0.073***	0.001	-0.017***	0.067***	1	
EBITDA/revenue	0.009**	-0.011***	0.019***	0.010***	0.004	0.024***	0.177***	0.028***	-0.002	-0.004	-0.025***	-0.005	1

*** p<0.01, ** p<0.05, * p<0.1

Table 5: Explanatory variables used in predictive models. Simple logistic regressions only include some of the variables to avoid possible overfitting and collinearity.

ESG	Fundamentals & Market
Base ESG score	Size (SMB)
E, S and G scores	P/E (HML)
Pillars for each score (10 total)	CAPEX-to-revenue (CMA)
CO2 eq. intensity/revenue	EBITDA-to-revenue (RMW)
CO2 Scope 1 – direct	Past year return
CO2 Scope 2 – indirect	Revenue
CO2 Scope 3 – indirect, up and down-stream	Book-to-market ratio (HML)
	Debt-to-equity ratio
	Industry dummies (Sin-stock S proxy)
	Country dummies

3.3 Sample

The data is limited by the availability of ESG metrics, as discussed above and illustrated in figures 1 and 2. For the purposes of improving prediction accuracy, I have chosen to include all data available instead of using a subsample like S&P500 firms or even firms from developed markets only. Most relevant companies should be covered with the Refinitiv ESG coverage being market as currently covering 85% of global equity (LSEG, 2022).

The sample drops as forward-looking data about controversies limits the use of recent years, where the number of observations would be higher due to better coverage. That is to say, for a one-year forward model, the sample ranges from 2003 to 2023 future controversies and from 2002 to 2022 for predictor variables. As ESG coverage is required for both the predictor variables at $t=0$ and for the evaluated controversy year $t+1$, the most recent year is eliminated from the sample. From the initial total coverage of 89,040, the number of observations used in the model 1 of table 8 is 72,773.

The other two factors limiting data is missing fundamental data and perfect predictors for countries and industries, together leading to the drop of 3,300 observations in model 1 of table 8. Of this, 419 observations dropped were perfect predictors. Out of the sub-industries, Mortgage REITs (40204010), Hotel & Resort REITs (60103010) and Self-Storage REITs (60108020) were

eliminated due to lack of variation, along with 25 countries. The “country” with the most eliminated observations was Guernsey with 58 omissions.

The following frequency tables 5 and 6 are based on the sample of model 1 from table 8 with 72,773 observations. The alternative with a two-year-forward controversy outcome and models with more control variables will have less observations. Table 5 presents the geographical distribution of the firm-year observations based on company headquarters, and table 6 aggregates the 163 subindustries into 25 industry groups according to GICS classifications.

Table 6: Distribution firm-year observations by country

US	21,377	29.4 %
Japan	6,598	9.1 %
China	4,696	6.5 %
Great Britain	4,453	6.1 %
Canada	3,464	4.8 %
Australia	3,194	4.4 %
Germany	1,809	2.5 %
France	1,697	2.3 %
Hong Kong	1,670	2.3 %
Taiwan	1,665	2.3 %
Sweden	1,567	2.2 %
Others (57 pcs)	20,583	28.3 %
Europe total	17,245	23.7 %
TOTAL	72,773	100.0%

Table 7: Distribution of industries into GICS industry groups

GICS industry group code and name		Firm-year observations	%
2010	Capital Goods	8,222	11.3 %
1510	Materials	7,195	9.9 %

4010	Banks	5,342	7.3 %
4020	Financial Services	3,933	5.4 %
1010	Energy	3,781	5.2 %
3520	Pharmaceuticals, Biotechnology & Life Sciences	3,622	5.0 %
5510	Utilities	3,135	4.3 %
3020	Food, Beverage & Tobacco	2,997	4.1 %
4520	Technology Hardware & Equipment	2,698	3.7 %
2030	Transportation	2,677	3.7 %
3510	Health Care Equipment & Services	2,550	3.5 %
2550	Consumer Discretionary Distribution & Retail	2,544	3.5 %
6010	Equity Real Estate Investment Trusts (REITs)	2,423	3.3 %
2520	Consumer Durables & Apparel	2,338	3.2 %
5020	Media & Entertainment	2,288	3.1 %
4030	Insurance	2,272	3.1 %
4510	Software & Services	2,262	3.1 %
6020	Real Estate Management & Development	2,195	3.0 %
2530	Consumer Services	2,115	2.9 %
2020	Commercial & Professional Services	1,887	2.6 %
2510	Automobiles & Components	1,627	2.2 %
5010	Telecommunication Services	1,576	2.2 %
4530	Semiconductors & Semiconductor Equipment	1,355	1.9 %
3010	Consumer Staples Distribution & Retail	1,077	1.5 %
3030	Household & Personal Products	662	0.9 %
	Total	72,773	100.0%

The Refinitiv ESG coverage is not uniquely defined by an objective metric like firm size, but attributes such as index inclusions affect the decisions to extend coverage. The selection process is not entirely transparent, leaving room for selection bias. For example, companies, which believe themselves to be above competition in ESG metrics, would be more inclined to contribute to the data. LSEG in fact entice companies to “Showcase your firm's ESG data” by contributing to the data. It remains unclear whether the companies own actions to contribute influence the decision to extend ESG coverage to that company.

Second source of possible sample bias arises as the forward-looking ESG controversy metric will, by definition, have survivorship bias. Companies, which are delisted due to bankruptcy or privatisation, or which are otherwise no longer covered by Refinitiv, will not be present in the sample. In terms of predicting controversies, this could have a significant impact as an ESG controversy could be a factor in bankruptcies or other delistings. Data on delistings conditional on an ESG controversy (or even an impending controversy which is not publicly disclosed) is not readily available, and such analysis would require more complex methods such as machine reading text to identify these situations.

As discussed in the introduction, the objectivity of the ESG scores and the controversy score is also questionable. Firms attracting more societal attention and scrutiny will have more controversies revealed by the media. Large firms have a disproportionately large share of controversies. Similarly, companies from countries, where reporting is done in English, more represented. This causality arising from the firm's headquarters is of course purely speculative, but it would fit the narrative of increased global media attention leading to more controversies. Further study for differences between countries including Hofstede's (2001) cultural dimensions theory for comparison would be interesting.

4 Methodology overview

For examples for predictive methods, I looked at credit rating and default risk prediction methods. Predictive models have been widely used for assessing default risk and to derive credit scores. ESG controversies are similar in many ways, namely predicting a seldom occurring event in companies with no track record of such events.

A systematic review by Markov et al. (2022) of the modern methods to derive credit scores looked at studies published between 2016 and 2021. In their sample of 110 papers, logistic regressions were used in 70 papers, support vector machines (SVMs) were used in 53 studies and linear regressions in 5. More complex machine learning methods were presented to challenge these baseline models, but in this thesis, I have opted to use logistic models enhanced by an elastic net model selection algorithm for predicting. Markov et al. also note that SVMs have become such a standard practice for credit risk analysis, that newer models use it as a baseline in benchmarking.

The elastic net algorithm can be reduced to a linear support vector machine as demonstrated by Zhou et al. (2015). Under parallel parameters, the results are the same for both methods with the more modern SVMs providing a much more efficient solver, which can utilize parallel computation and GPU processing. SVMs, however, have not been implemented into STATA. For the dataset and models that I am using, the elastic net was sufficient, although somewhat time-consuming to compute. Current credit scoring literature is of course more focused on the state-of-the-art methods, where algorithm optimisation is more relevant considering the size of datasets.

The other more involved machine learning methods like neural networks or bagging and boosting methods had to be left out of scope due to the increase in complexity and computational intensity. While better models would increase the out-of-sample performance, they would not necessarily provide more insights into interactions between firm characteristics and controversies. Logistic regressions are common in finance literature and thus provide more intuitive and interpretable results for readers. The models presented in this thesis can be run on a home computer in under a day without the need for cluster computing or other similar methods. For further research, comparisons between sophisticated methods and richer data could provide interesting insights especially into combined interactions that the models presented here cannot reveal as variables are not explicitly interacted in the models.

The study on ESG controversies in the banking sector by Agnese et al. (2023) uses a generalized method of moments (GMM) model to avoid

endogeneity. Using lagged right-hand side variables could pose an issue in biased and inconsistent estimators if overlap were to occur. In my analysis, I will be using a two-year lag to avoid this issue. The predictive models from the elastic net selection process are optimized using cross-validation for out-of-sample performance, so the theoretical constraints are not as important, as long as the predictive data is all ex-ante information.

The specific models and methodology are explained further in subsequent sections.

5 Logistic regressions results

5.1 Methodology

Following Kaustia and Zhang (2023), due to the categorical nature of the outcome variable of future controversies either occurring or not, I estimate the first, baseline models with logistic regressions. As in credit default predictions and scoring models, logistic regressions serve as a good starting point providing a methodologically easy and intuitively understandable baseline.

I classify ESG controversies into binary dummy variables in two different categories. The “Any controversy” category is triggered when the ESG controversy score is below the perfect 100 with the indicated timeframe of either two years forward $t+2$ or one year forward $t+1$. The “Severe controversy” is when the score is below or equal to the theoretical median of 50. The logistic regression is well suited for binary regressions like this, and linear regression with the same specifications gave qualitatively similar results, although for brevity I have not included them.

The model for logistic regressions I use is as follows:

$$ESGCS_{i,t+2} = \beta * P_{i,t} + \beta * F_i + \varepsilon_{it} \quad (1)$$

where $ESGCS_t$ is a dummy variable for whether the ESG controversies score are below the threshold at forward time $t+2$ or $t+1$. Betas and stock-specific predictors are paired in the vectors β and P , respectively. Year, country of primary listing headquarters and GICS sub-industry fixed effects, denoted by the vector F , are used in all logistic models unless otherwise specified.

For logit regressions, I report z-values corresponding to the coefficients of the independent variables to serve as a measure of their statistical significance. The asterixis in turn correspond to two-tailed tests for significance i.e. z-values exceeding 1.96 receive two asterixis and can be considered statistically significant, suggesting a robust effect of the independent variable on the dependent variable.

The control variables used are Log_{10} market capitalisation for size, EBITDA-to-revenue ratio for profitability, past year returns for momentum and visibility as standing out to investors, both good and bad. I include two dummy variables for past returns, a high and (low) grouping for firm-year observations where the return was in the highest (lowest) quartile. The mean of the thresholds was 29.44% p.a. for the high category and -15.87% for the low category. This was to capture the effect of standing out and to help clean the effect of having marginally better (worse) performance around the median. I

also include a square of the past year return to assess the effect of extreme outliers in driving future controversies. This is to test for the hypothesis of visibility from extraordinary returns, as the squared returns will capture both negative and positive outliers.

The models are all formed with a binary outcome variable of a controversy happening in either one or two years ahead of the explanatory variables. A two-year lag on all explanatory variables was used to avoid bias issues arising from jointly determined variables on both sides of the regressions. However, the one-year forward-looking model is more interesting from a practical perspective: a shorter horizon is much more interesting for practitioners such as asset managers. Comparing the results, the choice does not have an economically significant effect on coefficients with only small changes in magnitude.

Robustness

As a robustness check, I repeat the first model of table 8 with different subsamples of the data based on geography, size and year. Tables I, II and III of the appendices report the results respectively:

The geography split was between US and Europe separately (regressions 1 and 2 of table I), as well as US and Europe together (regression 3) as opposed to the rest of the world (regression 4).

The size split follows the Refinitiv categorisation according to market capitalisation: stocks valued at 10 billion USD or over are classified as large and stocks below 2 billion USD are classified as small with mid-cap in between. The results are in table II.

For table III and year splits, I somewhat stretch the older grouping to retain a similar number of firm-year observations in each group. The splits are 2011 or older ($n = 16\ 151$), between 2012 to 2017 ($n = 23\ 360$) and 2018 to 2022 ($n = 32\ 204$).

5.2 Results

ESG scores, fundamentals and stock market variables

Tables 8, 9 and 10 present regressions on ESG scores, pillar scores and sub-pillar scores with controls respectively. I present results from two different thresholds and with both one- and two-year differences in the outcome compared to predictor variables. The results are robust to all the aforementioned alternation in the models. The pseudo R2 is moderate in all regression, but the statistical significance of the results is strong with ESG scores and pillar scores all being significant at a 1% level in the z-test.

As discussed in the hypothesis formation, the direction of the effect is somewhat contrary to expectations, that is, ESG scores and their subcomponents on all levels are positive, meaning a higher-level correlating to a higher chance of future controversies. The results from all three tables confirm hypothesis 1 and 2 with some exceptions. Overall, Refinitiv ESG scores predict higher ESG controversies in all specifications of table 7 with similar results in lower levels of aggregation in tables 8 and 9, albeit some sub-pillars remain statistically insignificant. The EBITDA-to-revenue ratio does not receive a significant loading with the entire sample. Yet in the most recent 2018 to 2022 and the mid-cap subsamples presented in the appendix, the ratio receives a small, negative loading, significant on a 1% level in z-tests.

As for the correlation with past returns, the linear past return receives a negative loading, whereas the high return dummy is positive, as is the squared return. The low category dummy is insignificant. Together these results point to rather complex form: stocks that perform moderately well will have less controversies, but stocks with exceptionally high returns or outliers in both directions will encounter more controversies.

In sub-pillars, the correlations are not as clear-cut as in pillar scores. Social pillar scores and corresponding sub-pillar scores receive the highest coefficients with consistent significance, except for the workforce sub-pillar. From the environmental pillar, emissions and environmental innovation receive significant loadings whereas resource use – a pillar with a 0.82 correlation with emissions sub-pillar – remains insignificant.

Interestingly, the governance sub-pillars are not consistently significant. They all receive significant loadings in some models, but none of them are significant in all of the different model specifications. The corporate social responsibility (CSR) strategy is most prevalent with significant loading in three of the four models. The management and shareholder pillars are somewhat more distant from other sub-pillars, as can be seen in the correlation matrix in table 3 of section 3.1.1. Here, they are mostly small in both magnitude and significance, yet inconsistent in when comparing outcomes of all controversies to significant controversies. Out of all pillars, one could expect that governance would be the most important in determining whether misconduct, improper risk management or other bad governance practices, which could lead to ESG controversies, are prevalent in the company.

Table 8: ESG scores and controls

Logistic regression of controversy dummies on ESG scores controlling for firm size with logarithmic market capitalisation, profitability with an EBITDA-to-revenue ratio, past year returns linearly, squared and with high and low category dummies if the firm-year observation at $t=0$ was in the highest (lowest) quartile. All regressions include fixed effects on countries, years and GICS sub-industries.

	(1) Any Controversy $t+2$	(2) Significant Controversy $t+2$	(3) Any Controversy $t+1$	(4) Significant Controversy $t+1$
ESG Score	0.0187*** (23.53)	0.0198*** (18.67)	0.0225*** (29.35)	0.0236*** (23.08)
Log market capitalisation	1.298*** (47.61)	1.057*** (30.46)	1.231*** (48.77)	0.961*** (29.98)
EBITDA-to-revenue	-0.000259 (-0.579)	0.0000675 (0.136)	0.0000641 (0.125)	0.00007.11 (0.141)
Ret $_{t-1}$	-0.00892*** (-9.860)	-0.0102*** (-8.795)	-0.0135*** (-16.45)	-0.0141*** (-13.50)
Ret $_{t-1}^2$	0.0000526*** (11.11)	0.000630*** (10.55)	0.000689*** (15.57)	0.000775*** (13.98)
High Ret $_{t-1}$ dummy	0.205*** (4.645)	0.227*** (3.971)	0.210*** (4.934)	0.310*** (5.594)
Low Ret $_{t-1}$ dummy	-0.0590 (-1.128)	0.000875 (0.0130)	-0.00422 (-0.0882)	0.101* (1.648)
Constant	-16.74*** (-26.31)	-13.60*** (-21.55)	-16.02*** (-26.86)	-12.58*** (-20.81)
Observations	72,773	71,852	82,122	80,683
Pseudo R2	0.281	0.246	0.281	0.244

z-statistics in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 9: ESG pillar scores

Logistic regression of controversy dummies on ESG pillar scores controlling for firm size with logarithmic market capitalisation. All regressions include fixed effects on countries, years and GICS sub-industries.

VARIABLES	(1) Any Controversy $t+2$	(2) Significant Controversy $t+2$	(3) Any Controversy $t+1$	(4) Significant Controversy $t+1$
Environmental Pillar	0.00541*** (7.198)	0.00567*** (5.691)	0.00631*** (8.753)	0.00651*** (6.801)
Social Pillar	0.0109*** (12.29)	0.0108*** (9.206)	0.0136*** (15.92)	0.0142*** (12.59)
Governance Pillar	0.00212*** (3.391)	0.00307*** (3.724)	0.00227*** (3.797)	0.00264*** (3.337)
Log market capitalisation	1.272*** (46.03)	1.037*** (29.40)	1.199*** (46.81)	0.929*** (28.51)
EBITDA-to-revenue	-0.000265 (-0.597)	0.0000621 (0.126)	0.0000600 (0.117)	0.0000681 (0.135)
Ret $t-1$	-0.00882*** (-9.748)	-0.0101*** (-8.740)	-0.0134*** (-16.24)	-0.0139*** (-13.34)
Ret $t-1^2$	0.000520*** (10.99)	0.000625*** (10.46)	0.000680*** (15.36)	0.000766*** (13.81)
High Ret $t-1$ dummy	0.205*** (4.646)	0.228*** (3.983)	0.208*** (4.867)	0.309*** (5.566)
Low Ret $t-1$ dummy	-0.0598 (-1.142)	0.000337 (0.00499)	-0.00344 (-0.0718)	0.102* (1.663)
Constant	-16.37*** (-25.60)	-13.36*** (-21.01)	-15.53*** (-25.91)	-12.20*** (-20.03)
Observations	72,766	71,845	82,115	80,676
Pseudo R2	0.281	0.246	0.282	0.245

z-statistics in parentheses
 *** p<0.01, ** p<0.05, * p<0.1

Table 10: ESG sub-pillar scores

Logistic regression of controversy dummies on ESG pillar score components controlling for firm size with logarithmic market capitalisation. All regressions include fixed effects on countries, years and GICS sub-industries.

VARIABLES	(1) Any Controversy $t+2$	(2) Significant Controversy $t+2$	(3) Any Controversy $t+1$	(4) Significant Controversy $t+1$
Resource Use Score	0.000795 (1.068)	0.00150 (1.545)	0.00115 (1.613)	0.00101 (1.077)
Emissions Score	0.00304*** (4.063)	0.00230** (2.328)	0.00327*** (4.566)	0.00350*** (3.689)
Environmental Innovation Score	0.00160*** (3.107)	0.00272*** (4.101)	0.00212*** (4.338)	0.00316*** (5.004)
Workforce Score	0.000884 (1.148)	0.00118 (1.134)	0.000153 (0.206)	-8.16e-05 (-0.0818)
Human Rights Score	0.00181*** (3.428)	0.00309*** (4.553)	0.00276*** (5.469)	0.00438*** (6.725)
Community Score	0.00489*** (8.083)	0.00546*** (6.771)	0.00590*** (10.13)	0.00674*** (8.641)
Product Responsibility Score	0.00219*** (4.417)	0.000843 (1.297)	0.00268*** (5.602)	0.00191*** (3.046)
Management Score	0.000790* (1.672)	0.00135** (2.162)	0.000597 (1.318)	0.000845 (1.416)
Shareholders Score	0.000269 (0.603)	0.00120** (2.044)	0.000613 (1.430)	0.00132** (2.333)
CSR Strategy Score	0.00252*** (4.017)	0.000952 (1.152)	0.00347*** (5.793)	0.00170** (2.144)
Log market capitalisation	1.252*** (45.02)	1.022*** (28.80)	1.175*** (45.54)	0.912*** (27.80)
EBITDA-to-revenue	-2.77e-05 (-0.624)	3.03e-06 (0.0618)	4.57e-06 (0.0898)	3.88e-06 (0.0774)
Ret $t-1$	-0.00873*** (-9.651)	-0.00997*** (-8.639)	-0.0133*** (-16.14)	-0.0138*** (-13.26)
Ret $t-1^2$	5.19e-05*** (10.96)	6.21e-05*** (10.40)	6.79e-05*** (15.34)	7.64e-05*** (13.75)
High Ret $t-1$ dummy	0.203*** (4.605)	0.222*** (3.886)	0.207*** (4.837)	0.306*** (5.502)
Low Ret $t-1$ dummy	-0.0598 (-1.141)	0.00347 (0.0513)	-0.00388 (-0.0809)	0.105* (1.712)
Constant	-16.14*** (-25.16)	-13.19*** (-20.59)	-15.21*** (-25.31)	-11.89*** (-19.42)
Observations	72,766	71,845	82,115	80,676
Pseudo R2	0.282	0.247	0.283	0.247

z-statistics in parentheses
 *** p<0.01, ** p<0.05, * p<0.1

Alternative ESG metrics

ESG scores are often criticized for being opaque or opinionated, and often rightfully so as per evidence presented in the introduction and literature review. For a more objective approach, I used simple proxies for environmental and societal factors. I could not identify a simple proxy for governance, with the common instruments like modified Jones model for discretionary accruals (Dechow, 1995) involving evaluating two-stage regressions for all companies. To limit the scope of this thesis and to keep the results practically relevant, I opted to not include discretionary accruals or other governance proxies.

The Environmental concern industry is a binary indicator including fossil fuel industries. The sin-stock category includes industries commonly regarded as “sinful”: oil & gas exploration, coal & consumable fuels, weapons, casinos & gambling, alcohol, and tobacco industries.

CO₂ eq. is a metric for all greenhouse gas emissions which has been standardised to CO₂ tonnes to account for the different characteristics different greenhouse gases have in terms of their potency to contribute to global warming.

Regressions for environmental and societal proxies are reported in table 11, and the results show positive coefficients for the proxies. The significance, however, varies. Interestingly, the consistently significant and relatively large coefficient for the emissions score sub-pillar does not translate to a significant loading for the CO₂ equivalent greenhouse gas emissions intensity. My hypothesis on more sustainable companies having less controversies does not hold in this regard or is at the least inconclusive: lower carbon intensity does not correlate with controversies.

Table 11: ESG pillar proxies

Logistic regression of controversy dummies on proxies for the environmental and societal pillars controlling for firm size with logarithmic market capitalisation. All regressions include fixed effects on countries and years.

VARIABLES	(1) Any Controversy $t+2$	(2) Any Controversy $t+2$
Sin-stock	0.202*** (2.753)	-0.000741 (-0.0136)
Scope 1&2 CO2 eq. tons/revenue	0.0583 (0.677)	
Environmental concern		0.527*** (11.81)
Log market capitalisation	1.785*** (54.23)	1.721*** (80.21)
EBITDA-to-revenue	-2.31e-05 (-0.516)	-1.83e-05 (-0.433)
Ret $t-1$	-0.00978*** (-7.362)	-0.0118*** (-13.52)
Ret $t-1^2$	6.08e-05*** (8.203)	6.67e-05*** (14.69)
High Ret $t-1$ dummy	0.187*** (3.069)	0.233*** (5.523)
Low Ret $t-1$ dummy	-0.00178 (-0.0248)	-0.0614 (-1.219)
Constant	-19.08*** (-12.01)	-20.23*** (-37.19)
Observations	30,008	73,595
Pseudo R2	0.257	0.231

z-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Industry and country fixed effects

The industry and country fixed effects of model 1 of table 8 are explored in tables 12 and 13 respectively.

Table 12: 10 best and worst industries

Results for the fixed effects coefficients from model 1 of table 8. That is, a higher coefficient represents a higher rate of controversies and vice-versa. For the best industries i.e. fewest controversies, REITs are excluded. Interactive Home Entertainment was originally 18th, but I chose to exclude the REITs occupying most of the fewest controversies leaderboard.

	GICS Sub-industry	Coefficient	z-statistic
20302010	Passenger Airlines	2.174 ***	(7.267)
25102010	Automobile Manufacturers	1.519 ***	(5.064)
25201010	Consumer Electronics	1.431 ***	(4.358)
10102010	Integrated Oil & Gas	0.898 ***	(2.933)
15104040	Precious Metals & Minerals	0.887 **	(2.359)
50102010	Wireless Telecommunication Services	0.853 ***	(2.827)
25101020	Tires & Rubber	0.843 **	(2.453)
20201080	Security & Alarm Services	0.789 **	(2.067)
40203030	Diversified Capital Markets	0.752 **	(2.094)
50202020	Interactive Home Entertainment	0.746 **	(2.244)
45203030	Technology Distributors	-2.824 ***	(-2.702)
25504030	Home Improvement Retail	-1.907 ***	(-4.232)
40301050	Reinsurance	-1.788 ***	(-3.959)
60201020	Real Estate Operating Companies	-1.272 ***	(-3.390)
45301010	Semiconductor Materials & Equipment	-1.266 ***	(-3.504)
35203010	Life Sciences Tools & Services	-1.207 ***	(-3.587)
15101040	Industrial Gases	-1.166 ***	(-2.743)
25501010	Distributors	-1.13 **	(-2.088)
20201060	Office Services & Supplies	-1.072 **	(-2.205)
25203030	Textiles	-1.06	(-1.003)
	Observations	72,773	
	Pseudo R2	0.281	

z-statistics in parentheses
 *** p<0.01, ** p<0.05, * p<0.1

Table 13: 20 worst countries

Results for the fixed effects coefficients from model 1 of table 8. That is, a higher coefficient represents a higher rate of controversies and vice-versa.

Coefficient	Coefficient	z-statistic
VG - British Virgin Islands	5.376 ***	(3.777)
AZ - Azerbaijan	5.173 ***	(3.642)
NG - Nigeria	3.62 ***	(2.767)
IN - India	2.619 ***	(5.947)
KY - Cayman Islands	2.457 ***	(4.330)
IM - Isle of Man	2.415 ***	(3.166)
MT - Malta	2.262 **	(2.551)
JO - Jordan	2.208 ***	(3.255)
PR - Puerto Rico	2.162 ***	(2.831)
IL - Israel	2.064 ***	(4.434)
JE - Jersey	1.999 ***	(3.221)
AU - Australia	1.936 ***	(4.397)
IE - Ireland	1.924 ***	(4.278)
IS - Iceland	1.915	(1.638)
ZA - South Africa	1.861 ***	(4.168)
GB - United Kingdom	1.852 ***	(4.223)
NZ - New Zealand	1.841 ***	(3.923)
US - United States	1.826 ***	(4.184)
CZ - Czech Republic	1.803 ***	(3.030)
RO - Romania	1.755 *	(1.900)
Observations	72,773	
Pseudo R2	0.281	

z-statistics in parentheses
 *** p<0.01, ** p<0.05, * p<0.1

The results on industries include some predictable outcomes, such as financial companies like REITs (Real Estate Investment Trusts) exhibiting low negative correlations with future controversies. They are omitted from the table, as 7 of the top 10 best sub-industries were different REITs, and I would rather present more interesting findings concisely. In line with hypothesis on visibility, consumer facing industries are well represented with the highest rate controversies. Integrated Oil & Gas is also present as I had hypothesised.

Similarly, the lowest controversy industries do include a lot of business-to-business or raw material and component industries. Some outliers arise, such as Precious Metals & Minerals in the high category. Although it would be a raw material industry, it is also somewhat commonly scrutinised publicly for improper ESG values due to association with direct, vast environmental impacts, improper working conditions, high emissions and so forth.

For countries, the hypothesis on reporting language and proximity bias seems to hold true, with countries where English is either a primary language (e.g. Nigeria, Malta, UK, South Africa), one of many official languages (e.g. India) or very prevalent in society without any official status (e.g. Israel). The English-speaking countries present are not close in physical distance, do not necessarily represent similar cultures, legal systems or otherwise share much in common. Different non-sovereign territories of the UK and US also receive a high loading. Outside of the English-speaking countries, Western-European countries are not present, apart from an insignificant loading for Iceland.

Generally, the results of logistic models can be already seen as red flags for practitioners to identify companies potentially at risk of controversy. However, with ESG scores correlating with increased future controversies, and sin-stocks also correlating positively, drawing clear conclusions can be rather difficult. Following each sub-industry out of 163 is also painstaking and practically not too useful. The following section on a more comprehensive predictive model would be easier for daily tracking once setup.

6 Elastic net regressions

With the elastic net regressions, I had two goals: to make the best possible model for out-of-sample predictions and to make an intuitively simple conclusion with only the most impactful variables in terms of maximizing out-of-sample accuracy. While the out-of-sample performance could be improved with more sophisticated methods, a longer sample or more explanatory variables like interaction variables, these results already show the most significant predictors in terms of standardised coefficients out of the variables explored in the literature review and data sections.

6.1 Methodology

For the predictive method, I use elastic net algorithm for model selection. Essentially the method chooses from a blend of LASSO (least absolute shrinkage and selection operator) and Ridge regressions by choosing a blending parameter alpha between 0 (Ridge model) and 1 (LASSO). Alpha determines the form of the penalisation function used in regularisation. Regularisation methods are used to deal with overfitting, allowing the use of a larger number of explanatory variables as well as variables which may be collinear. The elastic net methodology, benefits compared to ordinary least squares or plain LASSO regressions and other key characteristics are well documented by, for example, Emmert-Streib and Dehmer (2019).

The LASSO method and all elastic methods with an alpha above zero will limit the number of coefficients. The regularisation will remove the weakest and most collinear coefficients while optimizing out-of-sample performance. The elastic net method evaluates combinations of alpha and lambda to select a model with the best out-of-sample performance. In this case, I have specified ten-folds for the cross-validation algorithm to minimize out-of-sample deviance. In ten-fold cross-validation, the sample is divided into ten subsets and the model is tested ten times. In each fold, 9 of 10 subsets are used to train the model and the remaining subset is used to test and calculate the out-of-sample performance metrics. This is repeated so that each subset of the original data is used as the testing subset once as illustrated in figure 3.

The model itself is a logistic regression, following the same form as formula 1 of the previous section. However, as stated, it is possible to include a much wider range of predictors and even predictors with very high collinearity such as both ESG score and their subscores, or multiple metrics for firm size. Using country and industry variables in the elastic net model will result in a model with only the countries and industries, which have had a significant effect on the out-of-sample outcome.

Figure 3: Ten-fold cross-validation. The cross-validation is done for each value of alpha and lambda. That is, each point in the (α, λ) grid is rated based the average out-of-sample performance of tests from each fold.

	Data subsample:									
Fold:	1	2	3	4	5	6	7	8	9	10
1	Train									Test
2	Train								Test	Train
...										
9	Train	Test	Train							
10	Test	Train								

6.2 Results

The best performing model was selected based on tenfold cross-validation minimizing out-of-sample deviance ratio. The model was fitted with all the data available, except for years. The year dummies themselves could not be included in the same way as in the fixed effects logistic regressions because for the purposes of predicting, year trends are ex-post information. Incorporating trends from past years like the average amount of controversies could have been added, but for the purposes of keeping the model simple, I have chosen to not include them. The possible predictive power from time-trends in controversies could be interesting as the scrutiny and attention to ESG matters has certainly risen in the past years. For a forward-looking model, however, using past trends to predict future trend might not be too impactful or intuitive.

To find the best model, the selection algorithm used an input dataset with 37 different ESG, stock market and fundamental variables, 163 sub-industry dummies and 94 country dummies. After dropping perfect predictor dummies, i.e. countries and industries without any controversy observations, the model was left with 271 variables. Of these, 219 remain in the best model, after the regularisation has dropped the weakest explanatory variables. From the remaining 219, 26 are firm-level variables, and 136 industries as well as 57 countries remain.

Results for the elastic net model selection are presented in table 14. To arrive at the parameters tested here, I first used sparser alphas to figure out a range to test more specifically. The first iteration of models returned optimal alphas between 0.75 and 1 leading to this iteration.

The model is in fact rather robust to selection of alpha, as can be seen in table 14 deviance ratios. The final lambdas, that is, the lambda where the deviance ratios start to rise again after the minimum, mostly result in very similar performance for the tested alphas around the optimal choice. The choice of lambda is also robust to changes as can be seen in the lambda before and after the optimal choice. The cross-validation results are presented graphically in figure 4, where the allocation of lambda is clearly robust remaining rather linear for a large range around from 0.005 to 0.008.

The number of coefficients is also not a greatly determining factor – smaller lambdas will include more coefficients, but they will be incrementally smaller and less significant as lambdas decrease.

The out-of-sample performance for logistic regressions is measured using out-of-sample deviance ratios, which a measure of the model's predictive performance on data not used during training. It compares the generalized linear model deviance of the model's predictions to the deviance of a null model i.e., a model that only includes an intercept:

$$Dev. ratio = 1 - \frac{Deviance\ of\ the\ model}{Deviance\ of\ the\ null\ model}$$

All in all, the additional value on top of the logistic regressions is in predictive power arising from the regularisation. Additionally, the standardised coefficients allow for a meaningful comparison of a wide range of coefficients in one model.

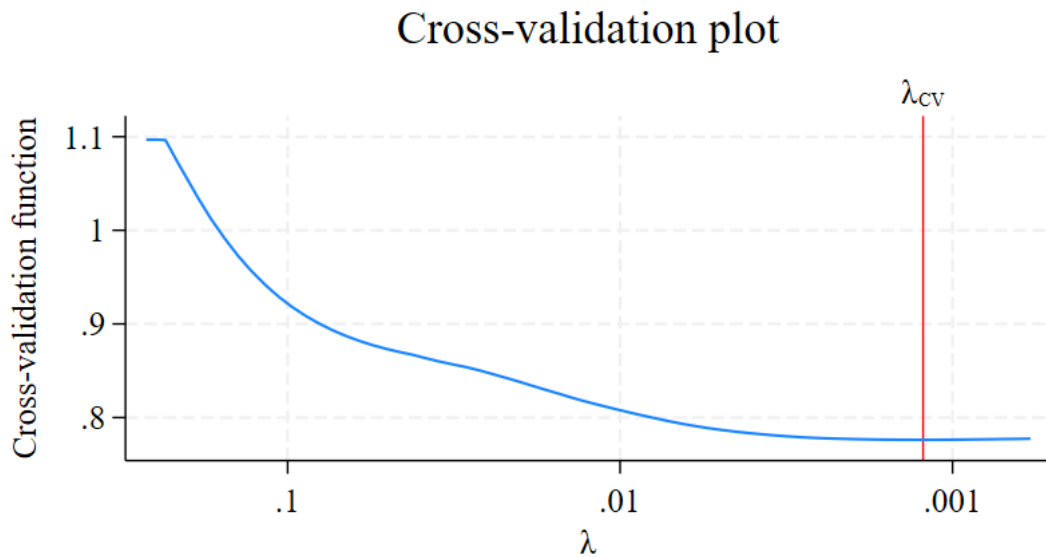
Table 14: Elastic net selection results

The elastic net model selection was run with different alphas in 0.1 increments. The optimal model is denoted with * at alpha 0.8 lambda 0.00123 16,832 observations, 271 covariates with 10-fold cross-validation.

Alpha	ID	Description	Lambda	No. of nonzero coef.	Out-of-sample dev. ratio	CV mean deviance
1	1	first lambda	0.26628	0	0.0001	1.096824
	72	last lambda	0.00044	247	0.2910	0.777590
0.9	73	first lambda	0.26628	0	0.0001	1.096824
	143	last lambda	0.00048	248	0.2910	0.777593
0.8	144	first lambda	0.26628	0	0.0001	1.096824
	203	lambda before	0.00135	213	0.2922	0.776219
	*	selected lambda	0.00123	219	0.2922	0.776187
	205	lambda after	0.00112	221	0.2922	0.776227
	212	last lambda	0.00058	246	0.2911	0.777396
0.7	213	first lambda	0.26628	0	0.0003	1.096410
	280	last lambda	0.00064	246	0.2911	0.777456

* Alpha and lambda selected by cross-validation.

Figure 4: Cross-validation plot showing the value of the CV function with regards to lambda at the optimal alpha of 0.8.



$\alpha_{cv} = .8$ is the cross-validation minimum α .
 $\lambda_{cv} = .0012$ is the cross-validation minimum λ ; # coefficients = 219.

Table 15: Strongest 36 coefficients.

Coefficients are ranked by based on standardised coefficients. Panel A reports the coefficients with the highest positive coefficients and panel B reports the lowest negative coefficients. Countries and industries refer to respective dummy-variables.

Panel A: most controversies			
Variable	Estimation coefficients	Standardised coefficients	Post-selection coefficients
Log revenue	1.258	0.882	1.278
Log total assets	0.378	0.297	0.325
Great Britain	0.714	0.217	0.805
t-1 total return	0.004	0.212	0.005
Log capital expenditure	0.230	0.189	0.226
India	1.020	0.159	1.073
Australia	0.795	0.139	0.888
Pharmaceuticals	0.973	0.138	1.128
South Africa	0.774	0.134	0.774
US	0.336	0.127	0.364
Common equity total	0.000	0.114	0.000
Automobile Manufacturers	1.026	0.109	1.188
Log market capitalisation	0.149	0.101	0.045
Passenger Airlines	1.286	0.098	1.362
Consumer Electronics	1.470	0.094	1.726
Health Care Equipment	0.882	0.086	1.017
Panel B: least controversies			
Past year FY0 return	-0.004	-0.374	-0.005
Japan	-0.893	-0.269	-1.003
Taiwan	-1.104	-0.183	-1.301
China	-1.365	-0.153	-1.633
Life & Health Insurance	-1.305	-0.145	-1.463
Diversified REITs	-1.497	-0.138	-2.250
Hongkong	-0.994	-0.123	-1.172
Office REITs	-1.360	-0.113	-2.417
Industrial REITs	-2.014	-0.112	-16.855
Multi-Sector Holdings	-2.153	-0.110	-17.093
Human Resource & Employment Services	-1.292	-0.105	-1.567
Real Estate Development	-1.246	-0.102	-15.546
Thailand	-0.862	-0.102	-1.088
Multi-line Insurance	-0.997	-0.098	-1.114
Retail REITs	-0.974	-0.097	-1.359
Turkey	-0.811	-0.087	-1.116
Indonesia	-1.607	-0.086	-25.032

Diversified Real Estate Activities	-1.021	-0.084	-1.340
Gas Utilities	-0.964	-0.080	-1.147
Electronic Manufacturing Services	-1.516	-0.079	-1.838

7 Conclusion

This thesis has demonstrated significant correlations between ESG metrics and future ESG controversies. Across all levels of aggregation, including overall ESG scores, pillar scores, and most sub-pillar scores, higher ESG performance is associated with a greater likelihood of future controversies. This counterintuitive finding suggests that confounding factors increased visibility or managerial opportunism might drive higher-scoring firms to encounter more controversies.

The logistic regression models reveal that higher Refinitiv ESG scores consistently correlate with an increased probability of future controversies. This relationship holds for overall ESG scores, E, S, and G pillar scores, and most sub-pillar scores. The models control for various factors, including firm size, profitability, past stock returns, industry, country, and yearly variations. Common “sin-stock” industries exhibit higher rates of future controversies, confirming that industry-specific differences in visibility, level of scrutiny faced, or unobservable ESG risks drive controversies.

Further, the visibility hypothesis is supported by the data, with firm size, extreme stock performance, and high-attention industries being significant predictors of controversies. Profitability, measured by EBITDA margins, is not a significant predictor, whereas past stock returns clearly are. English-speaking countries face more controversies than others, but Western European countries do not exhibit significant excess controversies, contrary to the initial hypothesis.

The elastic net method achieves an out-of-sample deviation ratio of 0.29 with ten-fold cross-validation, indicating the presence of predictability in future controversies using data mining methods. The most significant predictors identified are firm fundamentals, such as size (assets and revenue) and past return patterns. Industry and country-specific factors also play crucial roles, with large Asian economies (excluding India) predicting fewer controversies, while English-speaking countries, including Great Britain, India, Australia, South Africa, and the US, predict more.

The contribution of this thesis is two-fold. First, I show that the Refinitiv ESG scores are not a reliable measure of the ESG riskiness of a company. The results suggest that Refinitiv ESG controversies scores are biased in measurement, overly focusing on English-speaking countries, bigger firms attracting attention due to extraordinary stock returns of returns and firms belonging to visible sin-stock industries. These findings are yet another addition to evidence against the reliability of ESG ratings.

Second, I provide evidence that ESG controversies are not random but are predictable using firm characteristics such as ESG scores, fundamentals, and stock market performance indicators. My predictive model identifies the most significant factors correlated with future controversies, offering practical insights for practitioners. The ability to predict ESG controversies could significantly enhance portfolio management and risk management, providing a competitive advantage through the early identification of red flags.

For practitioners, implementing predictive models based on these findings can be highly impactful. Analysts and portfolio managers can leverage cloud-based predictive modelling services, which require minimal technical expertise and no proprietary computing power, making the implementation accessible and cost-effective. Regarding common ESG investing practices, the results speak for exclusionary screening based on industries, as different industries do face more controversies. However, screening or weighting portfolios based on ESG scores appears counterproductive, with higher scoring companies facing more controversies. For speculative trading, more frequent data would improve the possibilities to achieve alpha, as yearly ESG scores are perhaps not the most reliable to justify risky short sales or derivatives, for example.

This thesis opens several avenues for future research. With more computing power (GPU acceleration, cluster computing), more sophisticated predictive models could be developed using a broader range of variables and more advanced techniques such as deep learning, bagging, boosting, or neural networks. As the dataset continues to grow in both coverage and historical length, opportunities for more accurate predictions will expand. Future research could also explore interactions between various factors and incorporate additional data sources, such as market or analyst data, to uncover more nuanced relationships.

In conclusion, while the models presented in this thesis are relatively modest compared to modern data science capabilities, they still exhibit economically and statistically significant predictability. These findings underscore the potential for more advanced methodologies to further enhance the prediction and management of ESG risks, ultimately contributing to more sustainable and informed investment practices.

References

- Agnese, P., Battaglia, F., Busato, F., Taddeo, S., 2023. ESG controversies and governance: Evidence from the banking industry. *Finance Research Letters* 53, 103397. DOI: 10.1016/j.frl.2022.103397
- Altunbaş, Y., Thornton, J., & Uymaz, Y., 2018. CEO tenure and corporate misconduct: Evidence from US banks. *Finance Research Letters* 26, 1-8. DOI: 10.1016/j.frl.2017.11.003
- Aouadi, A., Marsat, S., 2018. Do ESG Controversies Matter for Firm Value? Evidence from International Data. *Journal of Business Ethics* 151, 1027–1047. DOI: 10.1007/s10551-016-3213-8
- Bloomberg, 2023. Global ESG assets predicted to hit \$40 trillion by 2030 despite challenging environment, forecasts Bloomberg Intelligence. [online] Available at: <https://www.bloomberg.com/company/press/global-esg-assets-predicted-to-hit-40-trillion-by-2030-despite-challenging-environment-forecasts-bloomberg-intelligence> [Accessed 9 June 2024].
- Berg, F., Fabisik, K. and Sautner, Z., 2021. Is history repeating itself? The (unpredictable past of ESG ratings). *European Corporate Governance Institute – Finance Working Paper* 708/2020, 1-59. <https://ssrn.com/abstract=3722087>
- Berg, F., Kölbel, J.F. and Rigobon, R., 2019. Aggregate Confusion: The Divergence of ESG Ratings. *Forthcoming Review of Finance*. <https://ssrn.com/abstract=3438533>
- Capelle-Blancard, G., Petit, A., 2019. Every Little Helps? ESG News and Stock Market Reaction. *Journal of Business Ethics* 157:2, 543-565. DOI: 10.1007/s10551-017-3667-3
- Carhart, M., 1997. On Persistence in Mutual Fund Performance. *Journal of Finance*, 52:1. DOI: 10.1111/j.1540-6261.1997.tb03808.x
- Chatterji, A. K., Durand, R., Levine, D. I., and Touboul, S., 2016. Do ratings of firms converge? Implications for managers, investors and strategy researchers, *Strategic Management Journal* 37, 1597–1614. DOI: 10.1002/smj.2407
- Chen, D., Chen, Y., Li, O. Z., & Ni, C., 2018. Foreign residency rights and corporate fraud. *Journal of Corporate Finance* 51, 142-163. DOI: 10.1016/j.jcorpfin.2018.05.004
- Cui, B., Docherty, P., 2020. Stock Price Overreaction to ESG Controversies. Working paper. <https://ssrn.com/abstract=3559915>

- DasGupta, R., 2022. Financial performance shortfall, ESG controversies, and ESG performance: Evidence from firms around the world. *Finance Research Letters*, 46B. DOI: 10.1016/j.frl.2021.102487
- De Franco, Carmine, 2020. ESG Controversies and Their Impact on Performance. *The Journal of Investing*, ESG Special Issue 29:2, 33-45. DOI: 10.3905/joi.2019.1.106
- Dechow, P.M., Sloan, R.G., & Sweeney, A.P., 1995. Detecting earnings management. *The Accounting Review*, 70(2), 193-225. DOI: 10.2307/2491047
- Dorflleitner, G., Kreuzer, C., Sparrer, C., 2020. ESG controversies and controversial ESG: about silent saints and small sinners. *Journal of Asset Management* 21, 393–412. DOI: 10.1057/s41260-020-00178-x
- Drempetic, S., Klein, C. and Zwergel, B., 2020. The influence of firm size on the ESG score: Corporate sustainability ratings under review. *Journal of Business Ethics* 167, 333-360. DOI: 10.1007/s10551-019-04164-1
- Dupont, Q., Karpoff, J.M., 2020. The Trust Triangle: Laws, Reputation, and Culture in Empirical Finance Research. *Journal of Business Ethics* 163, 217–238. DOI: 10.1007/s10551-019-04229-1
- Emmert-Streib, F. and Dehmer, M., 2019. High-Dimensional LASSO-Based Computational Regression Models: Regularisation, Shrinkage, and Selection. *Machine Learning & Knowledge Extraction* 1, 359-383. DOI: 10.3390/make1010021
- Fama, F., and French, K., 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116:1, 1-22. DOI: 10.1016/j.jfineco.2014.10.010
- Ferrés, D., Marcet, F., 2021. Corporate social responsibility and corporate misconduct. *Journal of Banking & Finance* 127, 106079,. DOI: 10.1016/j.jbankfin.2021.106079
- Gelman, M., Khan, Z., Shoham, A., & Tarba, S. Y., 2021. Does local competition and firm market power affect investment adviser misconduct? *Journal of Corporate Finance* 66, 101810. DOI: 10.1016/j.jcorpfin.2020.101810
- Godfrey, P., 2005 The Relationship between Corporate Philanthropy and Shareholder Wealth: A Risk Management Perspective. *The Academy of Management Review* 30:4, 777-798. DOI: 10.5465/amr.2005.18378878
- Hofstede, G., 2001. Culture's Recent Consequences: Using Dimension Scores in Theory and Research. *International Journal of Cross Cultural Management* 1:1, 11-17. DOI: 10.1177/147059580111002
- Karpoff, J.M., 2021. The future of financial fraud, *Journal of Corporate Finance* 66, 101694. DOI: 10.1016/j.jcorpfin.2020.101694

- Kaustia, M. and Zhang, Y., 2023. Are better ESG companies involved in more controversies? Working Paper, 12th International Research Meeting in Business and Management -conference.
- Klein, B., Leffler, K., 1981. The Role of Market Forces in Assuring Contractual Performance. *Journal of Political Economy* 89:4, 615-641. DOI: 10.1086/260996
- Krüger, P., 2015. Corporate goodness and shareholder wealth. *Journal of Financial Economics* 115:2, 304-329. DOI: 10.1016/j.jfineco.2014.09.008
- Li, J., Haider, Z.A., Jin, X., Yuan, W., 2019. Corporate controversy, social responsibility and market performance: International evidence. *Journal of International Financial Markets, Institutions and Money* 60, 1-18. DOI: 10.1016/j.intfin.2018.11.013
- Liu, C., 2018. Are women greener? Corporate gender diversity and environmental violations. *Journal of Corporate Finance* 52, 118-142. DOI: 10.1016/j.jcorpfin.2018.08.004
- Liu, X., 2016. Corruption culture and corporate misconduct. *Journal of Financial Economics* 122:2, 307-327. DOI: 10.1016/j.jfineco.2016.06.005
- LSEG, 2022. Environmental, social and governance scores from Refinitiv. May 2022. Available at: https://www.lseg.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf [Accessed 23 June 2024].
- Markov, A., Seleznyova, Z., Lapshin, V., 2022. Credit scoring methods: Latest trends and points to consider. *The Journal of Finance and Data Science* 8, 180-201. DOI: 10.1016/j.jfds.2022.07.002
- Neville, F., Byron, K., Post, C., & Ward, A., 2019. Board independence and corporate misconduct: A cross-national meta-analysis. *Journal of Management* 45:6, 2538-2569. DOI: 10.1177/0149206318801999
- Pelozo, J., 2006. Using Corporate Social Responsibility as Insurance for Financial Performance. *California Management Review* 48:2, 52-72. DOI: 10.2307/41166338
- Shiu, Y.M., Yang, S.L., 2017. Does Engagement in Corporate Social Responsibility Provide Strategic Insurance-like Effects? *Strategic Management Journal* 38:2, 455-470. DOI: 10.1002/smj.2494
- Tversky, A., Kahneman, D., 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5, 297-323. DOI: 10.1007/BF00122574

Velazquez, G., Oliver, J., 2023. Measuring ESG Risk: ESG Controversies Lead to a 2% to 5% Stock Underperformance after Six Months. Clarity AI. Available at: <https://clarity.ai/research-and-insights/measuring-esg-risk-esg-controversies-lead-to-a-2-to-5-stock-underperformance-after-six-months/> [accessed 30.10.2023]

Zhou, Q., Chen, W., Song, S., Gardner, J., Weinberger, K., & Chen, Y., 2015. A Reduction of the Elastic Net to Support Vector Machines with an Application to GPU Computing. Proceedings of the AAAI Conference on Artificial Intelligence 29:1. DOI: 10.1609/aaai.v29i1.9625

Appendix

Table I: Model 1 from table 8 with samples split by country of headquarters.

VARIABLES	(1) US Any Contro- versy $t+2$	(2) Europe Any Contro- versy $t+2$	(3) US & Europe Any Contro- versy $t+2$	(4) Rest of the world Any Contro- versy $t+2$
ESG Score	0.0272*** (15.17)	0.0125*** (8.537)	0.0178*** (16.42)	0.0189*** (15.08)
Log market capitalisation	1.403*** (23.68)	1.270*** (27.08)	1.344*** (38.13)	1.307*** (27.88)
EBITDA-to- revenue	1.403*** (23.68)	1.270*** (27.08)	1.344*** (38.13)	1.307*** (27.88)
Ret $_{t-1}$	-0.00928*** (-5.136)	-0.00922*** (-5.734)	-0.00931*** (-7.918)	-0.00875*** (-6.020)
Ret $_{t-1}^2$	4.10e-05*** (3.850)	5.22e-05*** (6.408)	5.13e-05*** (8.271)	5.39e-05*** (7.194)
High Ret $_{t-1}$ dummy	0.202** (2.362)	0.263*** (3.479)	0.225*** (4.064)	0.181** (2.420)
Low Ret $_{t-1}$ dummy	-0.0747 (-0.709)	-0.166* (-1.690)	-0.117* (-1.654)	0.0230 (0.288)
Constant	-17.28*** (-17.07)	-15.06*** (-18.43)	-16.43*** (-26.86)	-18.50*** (-25.05)
Observations	16,640	21,290	38,594	33,506
Pseudo R2	0.324	0.299	0.295	0.258

Table II: Model 1 from table 8 with samples split by stock market capitalisation.

>10 billion USD = Large

2-10 billion USD = Mid

< 2 billion USD = Small

VARIABLES	(1)	(2)	(3)
	Large	Mid	Small
	Any Controversy	Any Controversy	Any Controversy
	t+2	t+2	t+2
ESG Score	0.0163*** (12.05)	0.0178*** (14.05)	0.0213*** (10.68)
Log market capitalisation	1.994*** (28.36)	1.344*** (12.43)	0.361*** (4.787)
EBITDA-to-revenue	-1.40e-05 (-0.256)	-0.00114*** (-2.956)	0.000123 (0.232)
Ret _{t-1}	-0.00612*** (-3.567)	-0.00641*** (-4.238)	-0.00787*** (-4.739)
Ret _{t-1} ²	4.10e-05*** (3.850)	5.22e-05*** (6.408)	5.13e-05*** (8.271)
High Ret _{t-1} dummy	0.121* (1.699)	0.131* (1.790)	0.240** (2.080)
Low Ret _{t-1} dummy	-0.0191 (-0.204)	0.0333 (0.404)	-0.0831 (-0.755)
Constant	-23.73*** (-21.41)	-14.71*** (-12.30)	-8.905*** (-10.07)
Observations	16,640	21,290	38,594
Pseudo R2	0.324	0.299	0.295

Table III: Model 1 from table 8 with samples split by years.

VARIABLES	(1)	(2)	(3)
	2002-2011	2012-2017	2018-2022
	Any Controversy	Any Controversy	Any Controversy
	t+2	t+2	t+2
ESG Score	0.0215*** (14.53)	0.0178*** (12.91)	0.0151*** (10.47)
Log market capitalisation	1.647*** (27.49)	1.524*** (29.91)	1.093*** (25.99)
EBITDA-to-reve- nue	4.93e-07 (0.00931)	-0.000123 (-0.224)	-0.000902** (-2.342)
Ret _{t-1}	-0.0106*** (-6.603)	-0.0120*** (-7.306)	-0.00632*** (-3.995)
Ret _{t-1} ²	6.26e-05*** (7.141)	7.29e-05*** (7.839)	5.13e-05*** (8.271)
High Ret _{t-1} dummy	0.164** (2.066)	0.264*** (3.392)	0.263*** (3.358)
Low Ret _{t-1} dummy	-0.119 (-1.225)	0.0219 (0.246)	-0.141 (-1.514)
Constant	-18.95*** (-14.71)	-18.76*** (-18.16)	-14.94*** (-15.17)
Observations	16,640	21,290	38,594
Pseudo R2	0.324	0.299	0.295