Green Data Science Group Project - Examining the Relationship between Asset Utilisation Efficiency and Carbon Emissions using Asset Turnover as a Proxy

June 23, 2023

UCD Michael Smurfit Graduate Business School

Table 1: Student List

Student Name	**Student Number**
Umut Yasar	22205176
Evan Mc Garry	22206023
Sheethal Mohandas	22201844
Khuong Nguyen	22201697
Malavika Venkatesh	22201920

Word Count: 1996

Table of Contents

1 Executive Summary

The analysis conducted in this report revolves around *Asset Turnover* as this will act as the proxy for *Asset Utilisation Efficiency.*

- **Research Objective: Hypothesis Development**
  - The research conducted in literature review provides sufficient support to this report's claim that firms with higher "asset utilisation efficiency", which is proxied by "Asset Turnover", generally exhibit lower carbon emissions.
- **Research Design: Data Collection**
  - The Scope 1, Scope 2 Location-based and Scope 2 Market-based emissions data were scraped from the "Race To Zero" website in addition to the company names. 125 of the companies were randomly sampled.
  - The company names were then used with the Thomson Reuters Eikon Python API to collect the tickers for each of the 125 firms, which were necessary to collect Asset Turnover data as well as the control variables from Refinitiv Eikon. The emissions data was merged with the *lagged* financial data and structured as "Panel Data" to facilitate robust statistical analysis.
  - Summary statistics for all variables are included.
- **Research Findings: Statistical Analysis**
  - The dependent variables also include the sum of Scope 1 and Scope 2 Location-based emissions and the sum of Scope 1 and Scope 2 Market-based emissions.
  - A preliminary examination of the relationship between Asset Turnover and each of the dependent variables occurs through data visualisation.
  - The distributions of the dependent variables are examined to check for skewness and kurtosis.
  - The popular regression method for panel data "Panel OLS" (POLS) is used to examine the relationship between Asset Turnover and the dependent variables with fixed time-effects. Asset Turnover is statistically significant with a large negative growth rate coefficient which substantiates this report's hypothesis.
- **Research Limitations: Statistical Tests**
  - The time-effects are found to be jointly statistically *insignificant.* Alternative models, Pooled OLS and Ridge Regression, are proposed and subsequently tested in Appendix A.1 since PanelOLS with fixed time-effects is inappropriate for use.
  - This report does not consider entity-effects as, upon testing, the models had such weak explanatory power that entity effects were all that remained to explain the variance making it wholly unsuitable.
  - Statistical tests are performed on the Panel OLS Models which reveals further issues such as multicollinearity.
- **Research Implications: Importance of Asset Utilisation Efficiency**
  - The Panel OLS models exhibited strong *negative* growth rates which were statistically significant at the 10% significance level for 3 models. This indicates that the hypothesised relationship exists, is supported and can be explored further.
  - The explanatory power of the models was weak but this was forseen due to numerous issues found through examination of the data, statistical tests and evaluation of the models. With a larger data-set and more control variables, the relationship between Asset Utilisation Efficiency and Carbon Emissions could be leveraged into creating a model that allows companies to accurately predict how increased asset utilisation efficiency can improve their environmental performance which is correlated with improved financial performance.

2 Hypothesis Development

The reduction of carbon emissions can attract socially responsible investors (SRIs) who are seeking to maximise their profits while doing good. Accordingly, Asafo-Adjei et al. (2022) notes that this has initiated academic research into proxies of firm characteristics such as "Asset Turnover" and "Profitability" to examine the "contribution of CSR to firm performance". These studies have yielded numerous interesting conclusions which support the fundamental hypothesis of this report.

- For instance Dan et al. (2023) identifies **high total asset turnover as being one of the main factors creating differences in corporate financial performance**. Further it is noted that as *asset utilisation* becomes less efficient, the changes in carbon emissions becomes greater.
- This sentiment is echoed by Xu et al. (2021) who takes it one step further by stating that firms are advised to adopt green strategies which *directly enhance* Asset Turnover.
- Strong support for this relationship is proferred by Nurul Houqe et al. (2022) whereby firms that are dedicated to efficienct asset utilisation are involved in the optimisation of operational efficiency and waste reduction which leads to a by-product of *lower carbon emissions*.

Bukit et al. (2017) neatly summarises the benefit of this relationship as higher profitability being an indicator of efficient asset utilisation which suggests lower carbon emissions. The focus of this report is set firmly on the latter part of this statement.

- Despite this, profitability is a necessary control and will be accounted for by control variables such as "EBITDA Margin".
- As per Xu et al. (2021) one of the largest contributing factors to increases in Return on Equity (a key profitability metric) is Asset Turnover. This can be sector-dependent but it remains that Asset Turnover is a key metric in both profitability and carbon emissions, creating an inextricable link between the two.

This research provides sufficient support for this report's hypothesis which is formalised as:

- **Hypothesis 1 (H1)**: High-performing firms in terms of asset utilisation efficiency, using "Asset Turnover" as a proxy, generally exhibit lower carbon emissions.

3 Data Collection & Discussion

3.1 Sample Selection

The Race To Zero website has 500 members and the Data Explorer portion of the website lists each category of emissions from 2018-2021 for each company. This provides us with an entity, timeframe and observations which allows for structuring of panel data.

A Python script was created to scrape the company name and corresponding emissions data for each year across all emissions categories for all 500 companies. The scraped data was merged against a database that categorised each company by sector. Random sampling was performed on the data set to narrow the focus down to 125 of these companies.

A potential problem in this analysis would be the over-representation of one sector which would lead to *biased estimations* of population parameters. The random sampling resulted in the following:

Table 2: Proportion of Companies by Sector

Sector	Counts	Proportion
Industrials	26.0	21.14%
Consumer Discretionary	25.0	20.33%
Information Technology	15.0	12.20%
Consumer Staples	14.0	11.38%
Materials	13.0	10.57%
Health Care	11.0	8.94%
Utilities	9.0	7.32%
Communication Services	7.0	5.69%
Consumer Cyclicals	1.0	0.81%
Real Estate	1.0	0.81%
Technology	1.0	0.81%

There are no issues of representation present in the random sample. It also noteworthy that the "Consumer Non-Cyclicals" sector is excluded due to a severe lack of data for this sector which was only comprised of 2 companies, resulting in a random sample of 123 companies.

3.2 Data Collection

The Refinitiv Eikon Python API was used to extract the tickers for each company which were stored in a ".csv" file to facilitate the creation of a function in the Refinitiv Eikon Excel add-in that returns the requested financial data from Reuters' database.

The financial data collected represents the explanatory variables that will be used in this model. In total, 7 explanatory variables were retrieved from Refinitiv Eikon for each company and were *lagged* (ranging from 2017-2020). "Total Assets" and "Total Expenses" were used to calculate the "Expense Ratio" which replaced these 2 variables - resulting in 6 variables. The resulting data was converted to flattened panel data:

Table 3: Panel Data - Explanatory Variables

id (Ticker)	Dates	Sector	Asset Turnover	Current Ratio	EBITDA Margin, Percent	ESG Score	Quick Ratio	Expense Ratio
002555.SZ	2018-12-31	Communication Services	0.796177	2.60977	27.457111	27.347583	2.54271	0.509936

The financial data is now structured in the same format as the emissions data:

Table 4: Panel Data - Independent Variables

id (Ticker)	Dates	Sector	Scope 1	Scope 2 Location	Scope 2 Market	Scope 1 + 2 Location	Scope 1 + 2 Market
002555.SZ	2018-12-31	Communication Services	84240.038462	816850	275818.2	901090.038462	360058.238462

These two data-sets were merged to create the final panel data set. Any missing values were back-filled or forward-filled for each company where possible but if no data existed then the sector average was used.

3.3 Preliminary Statistical Analysis

The summary statistics of both explanatory and independent variables are provided below to get an idea of potential issues:

Table 5: Summary Statistics

Variables	**Mean**	**Median**	**Std. Dev.**	**Skewness**	**Kurtosis**
Asset Turnover (X1)	0.94	0.82	0.56	1.54	3.82
Current Ratio (X2)	1.40	1.25	0.72	2.78	10.69
EBITDA Margin (X3)	16.09	13.89	14.55	-1.27	8.87
ESG Score (X4)	67.41	69.88	15.43	-0.88	1.37
Quick Ratio (X5)	1.10	0.95	0.67	3.29	14.68
Expense ratio (X6)	0.83	0.73	0.54	1.71	4.98
Scope 1 (Y1)	2848245.64	129598.00	12926435.99	7.14	53.56
Scope 2 Location (Y2)	473861.72	167572.73	867759.12	5.20	40.62
Scope 2 Market (Y3)	378568.69	135267.00	763436.05	4.32	22.49
Scope 1 + 2 Location (Y4)	3322107.36	434144.02	13188124.04	6.99	51.80
Scope 1 + 2 Market (Y5)	3226814.33	281562.46	13275285.18	6.98	51.53

There is a minor issue with the scale of the dependent variables but a larger issue is present with the high positive skewness and excess kurtosis of the dependent variables. This suggests a log-normal distribution which will be analysed further in the next section.

4 Statistical Analysis

4.1 Data Visualisation

This sub-section will visually inspect the data to glean any necessary information before model creation.

4.1.1 Variable Analysis: Pre-Transformation

Given the high skewness and kurtosis of the emissions data, the distribution of this data is checked:

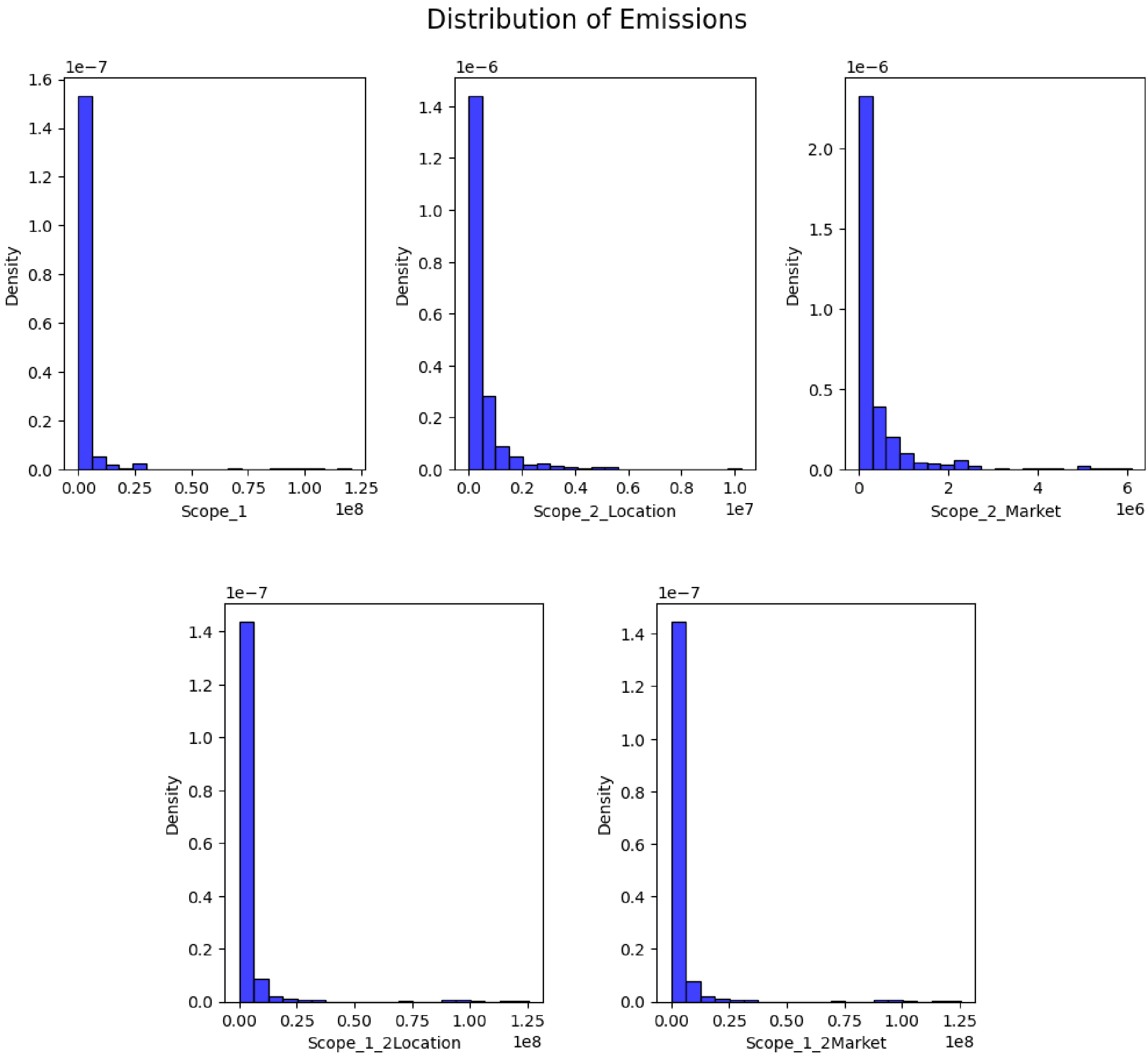

Figure 1: "Histograms" showing the Probability Density Function of each Dependent Variable.

These distributions exhibit characteristics of a log-normal distribution. A log-transformation is required to reduce the positive skewness. This transformation hinders the interpretability of regression results as the regression will become "Log *vs.* Level". Regression coefficients will now refer to increases or decreases in the growth rate (percentage) rather than units.

Before the transformation, it is prudent to examine the relationship between our dependent variables and Asset Turnover. This is done by "Sector" to also catch any impact a particular sector may have on emissions:

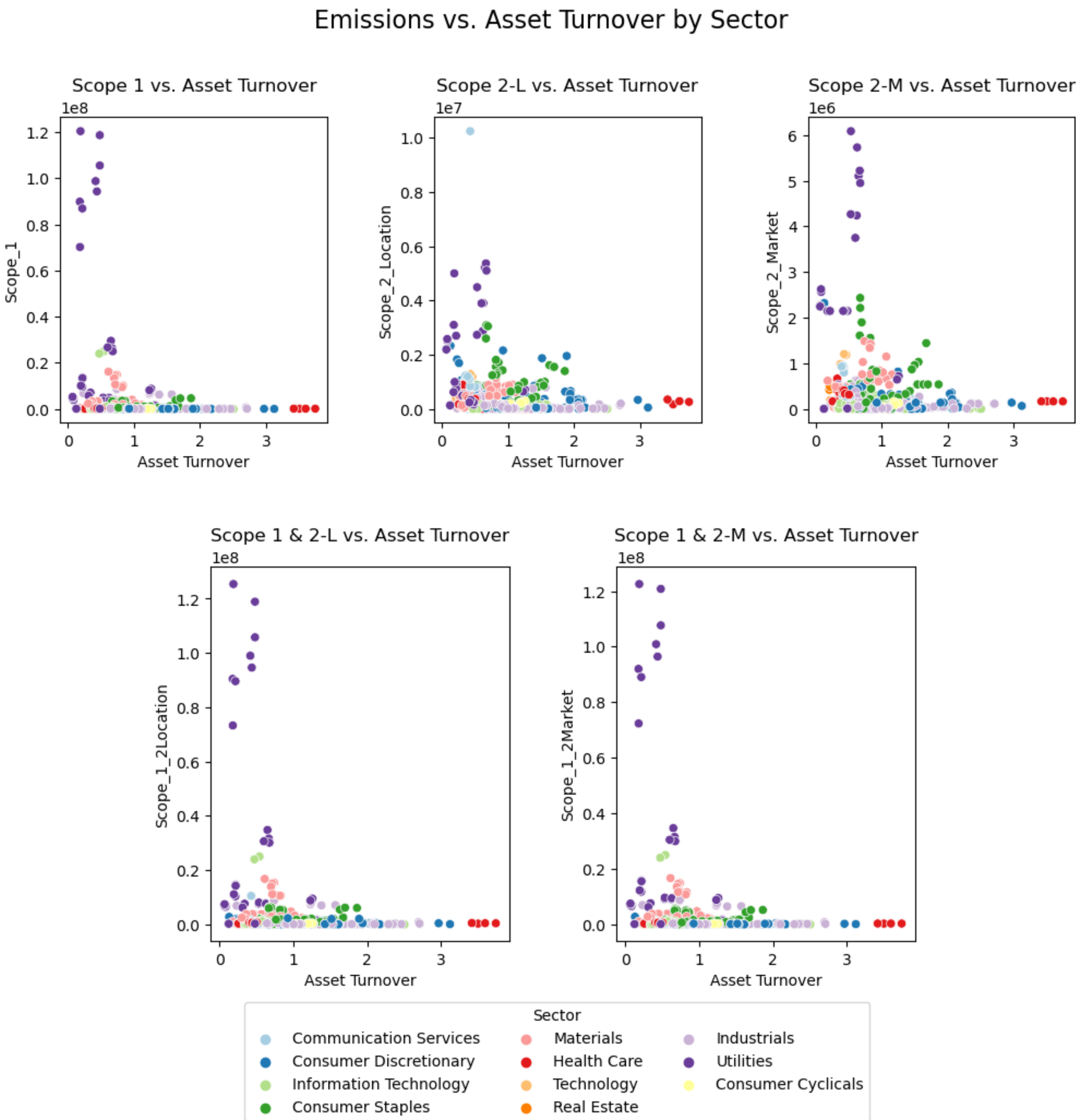

Figure 2: "Scatterplots" showing the Relationship between Asset Turnover and each Dependent Variable.

It is evident that emissions decrease as asset turnover increases but the scaling makes it difficult to assess. Interestingly, the "Utilities" sector exhibited the highest emissions across all categories. In order to **capture the impact of the Utilities sector** a **dummy variable was created** to categorise firms as being part of the Utilities sector or not. The categorisation of each sector with dummy variables is ill-advised as it would introduce high levels of multicollinearity into the model, rendering the results of the models unusable.

4.1.2 Variable Analysis: Post-Transformation

To reduce skewness and kurtosis a log-transformation was applied to the emissions data. This reduces the impact of outliers significantly and brings the distributions of the dependent variables closer to a normal distribution. This is particularly important for the construction of confidence intervals given that the sample being used is relatively small.

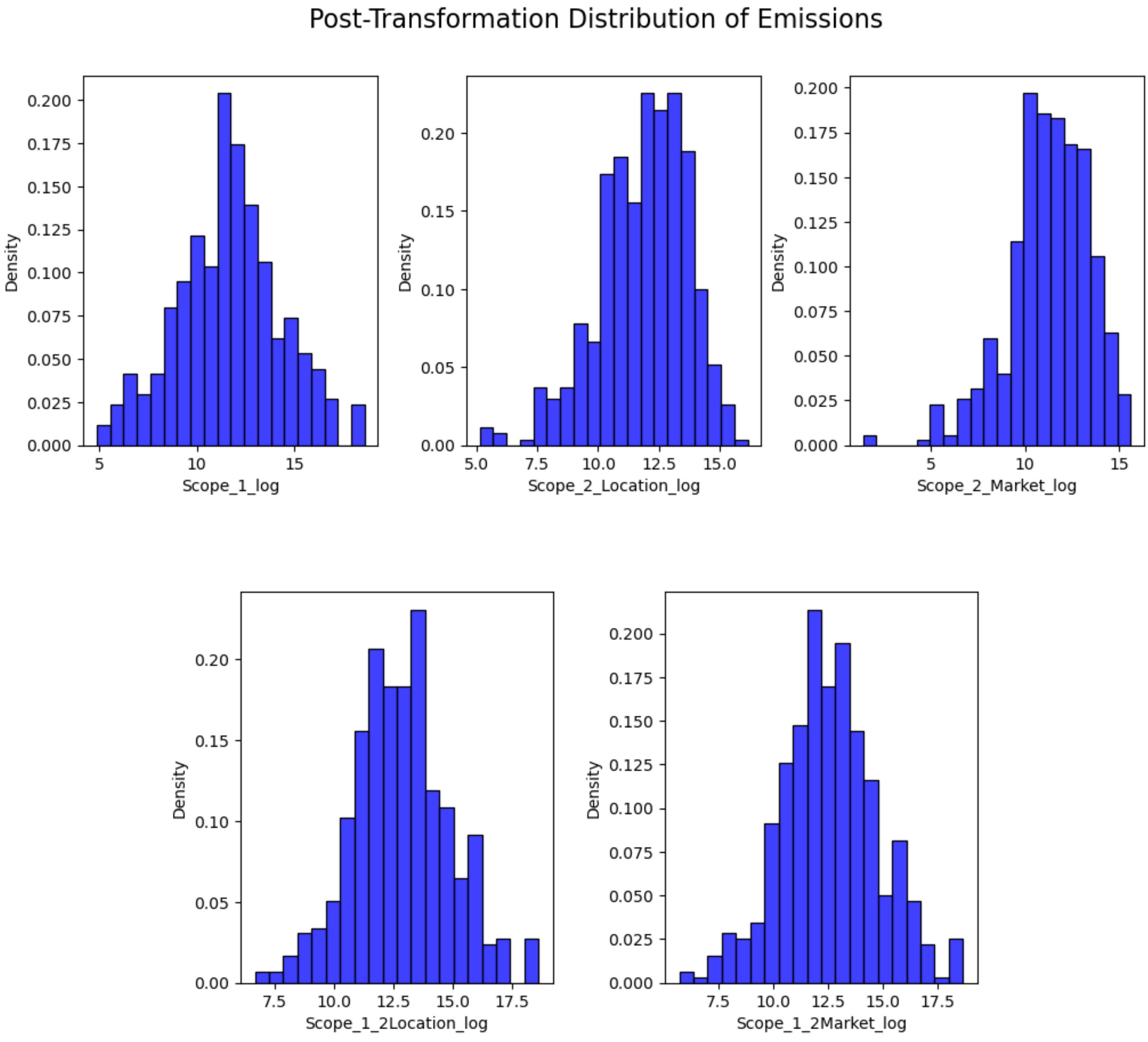

Figure 3: "Histograms" showing the Post-Transformation Probability Density Function of each Dependent Variable.

The resulting distributions appear to have lower degrees of kurtosis and skewness but some of them may not still be normal distributions. This should improve the standard errors of the models and give more accurate results.

Before discussing the panel regression models it is important to consider whether the hypothesised relationship holds when ignoring cross-sectional and time effects. Regression plots are created to visually interpret the relationship between Asset Turnover and emissions. This does not represent the significance of Asset Turnover in the Panel OLS models but if the fixed effects are found to be statistically insignificant then these plots offer insight into a potential linear relationship between the 2 variables:

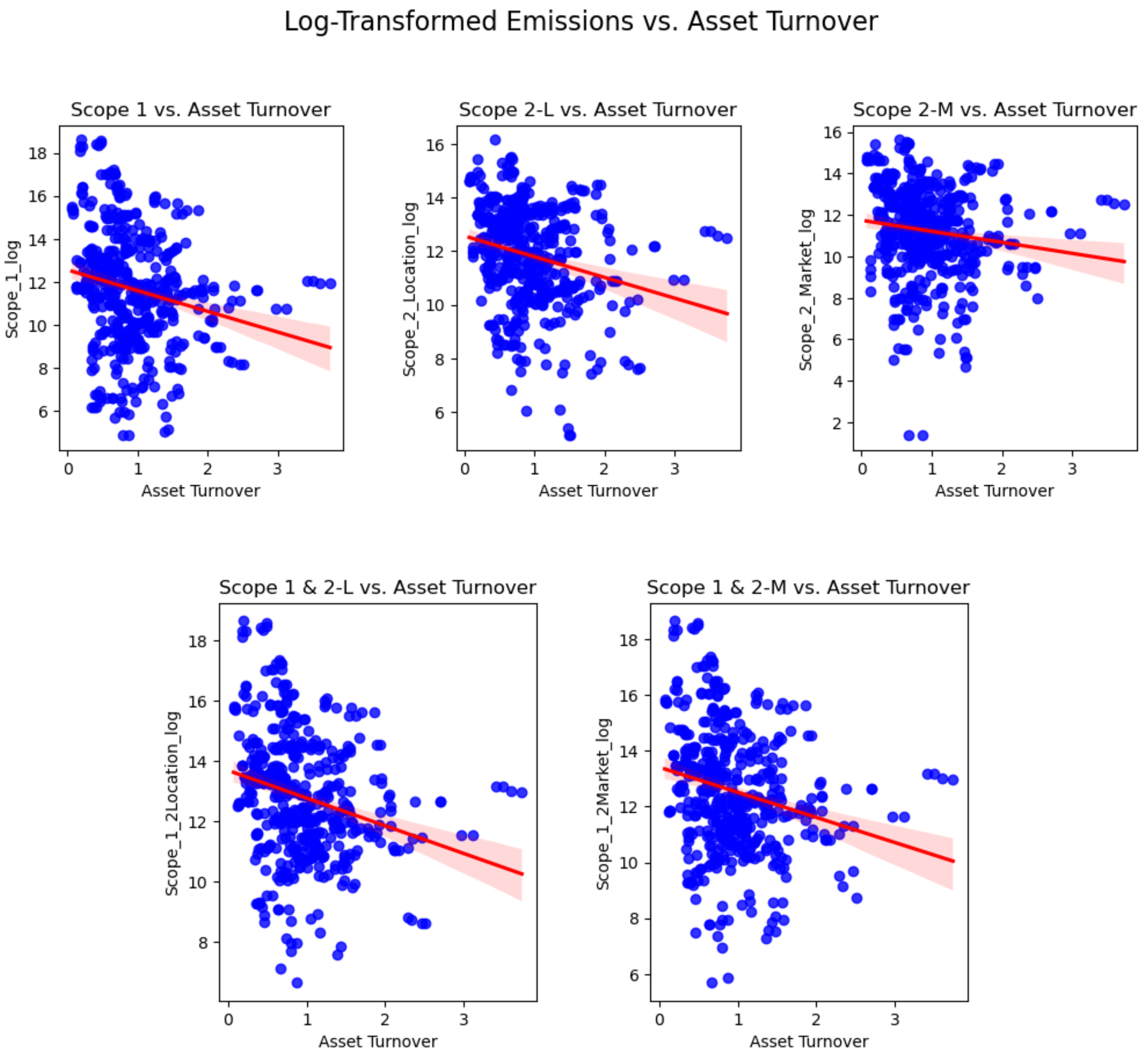

Figure 4: "Regression Plots" showing the Relationship between Log(Emissions) and Asset Turnover

The regression plots clearly show the existence of the hypothesised relationship and it appears to be a relatively strong linear relationship. Since we are using linear regression models, it is likely that Asset Turnover will have a high negative coefficient and be statistically significant.

4.2 Panel OLS Regression Models

4.2.1 Model Specification

The financial data has undergone manipulation throughout this report with additional variables being added. Therefore it is appropriate to discuss model specification with a table summarising the *explanatory variables* as well as their reason for inclusion:

Table 6: Model Specification

Explanatory Variables	Selection Reason
Asset Turnover (X1)	Key Variable**
Current Ratio (X2)	Liquidity Control
EBITDA Margin (X3)	Profitability Control
ESG Score (X4)	Performance Control
Quick Ratio (X5)	Liquidity Control
Expense Ratio (X6)	Profitability Control
Utilities (X7)	Sector Impact

This table lists the explanatory variables that are contained in the panel data and will act as the input for *every* model. With the explanatory variables for each model established, it is necessary to list the dependent variable used in each model:

Table 7: Dependent Variables

Dependent Variables	Model
log(*Scope 1 Emissions*) (Y1)	Model 1
log(*Scope 2 Location Emissions*) (Y2)	Model 2
log(*Scope 2 Market Emissions*) (Y3)	Model 3
log(*Scope 1 + Scope 2 Location Emissions*) (Y4)	Model 4
log(*Scope 1 + Scope 2 Market Emissions*) (Y5)	Model 5

The dependent variables are listed as the log of each category of emissions with the corresponding model. These tables establish clearly the explanatory variables and dependent variable being used in each model. It is crucial to remember at this point that these variables are structured as **panel data** as discussed previously.

4.2.2 Panel OLS Regression: Empirical Results

Panel data comprises observations for different cross-sections across time and in order to fit a linear regression model onto this data it was necessary to implement the Panel Ordinary Least Squares regression technique. The benefit of using this regression technique as opposed to a multiple linear regression is that it allows for the testing of time effects and entity effects (*i.e.* company) to see if these are important in explaining the variance of the dependent variable.

These 5 PanelOLS models were created and fitted to the panel data in Python. The table below compares the results of the first 3 regressions which are the standalone Scope 1 and 2 emissions:

Table 8: Model Comparison

	Model 1	**Model 2**	**Model 3**
Dep. Variable	Scope 1 Log	Scope 2 Location Log	Scope 2 Market Log
Estimator	PanelOLS	PanelOLS	PanelOLS
No. Observations	492	492	492
Cov. Est.	Clustered	Clustered	Clustered
R-squared	**0.2746**	0.2091	0.1728
R-Squared (Within)	-2.2866	-1.9404	-0.3947
R-Squared (Between)	0.2899	0.2344	0.2169
R-Squared (Overall)	0.2729	0.2050	0.1628
F-statistic	26.017	18.169	14.358
P-value (F-stat)	0.0000	0.0000	0.0000
=========================	=============	======================	===================
const	13.218***	13.004***	12.420***
	(15.449)	(28.141)	(20.907)
Asset Turnover	**-2.2568***	-1.5123	**-2.9179****
	(-1.6626)	(-1.5433)	(-2.5723)
Current Ratio	0.2026	-0.1076	-0.1526
	(0.3228)	(-0.1520)	(-0.2217)
EBITDA Margin, Percent	-0.0039	0.0184*	0.0148
	(-0.1568)	(1.9377)	(0.9179)
ESG Score	2.111e-10	1.492e-11	2.31e-10*
	(1.3399)	(0.1308)	(1.8159)
Quick Ratio	-1.1851*	-0.7236	-0.7455
	(-1.7299)	(-1.0071)	(-1.0117)
Expense Ratio	1.6093	1.0015	2.7185**
	(1.1507)	(0.9291)	(2.3966)
Utilities	**3.7988*****	0.9931	1.1993
	(4.8873)	(1.5929)	(1.4875)
===========================	===============	========================	=====================
Effects	Time	Time	Time

Asset Turnover was statistically significant at the 10% significance level in the first model and at the 5% significance level in the third. However it was statistically insignificant in the second model so the results will not be discussed.

- **Model 1:**
  - The Asset Turnover coefficient is approximately -2.26 meaning for a 1 unit increase in Asset Turnover there is an expected 226% decrease in Scope 1 Emissions at the 10% significance level. This seems unrealistic but a 1 unit increase in Asset Turnover is a large increase since it is a ratio making this result plausible.
  - The most interesting result is that the Utilities dummy variable is statistically significant at the 1% significance level with a coefficient of approximately 3.8. This means that if a firm is within the Utilities sector there is an expected 380% increase in Scope 1 Emissions. This result is plausible and was expected given the inclusion of the dummy variable as a result of the data visualisation process.
- **Model 3:**
  - The Asset Turnover coefficient is approximately -2.92 meaning for a 1 unit increase in Asset Turnover there is an expected 292% decrease in Scope 1 Emissions at the 5% significance level.

These models both have weak explanatory power with Model 1 being the stronger of the two with an R-squared of 27.46%. The statistical significance of Asset Turnover in the majority of models in addition to the highly negative coefficient lends strong support to this report's hypothesis. The remaining 2 models are now considered.

Table 9: Model Comparison

	Model 4	Model 5
Dep. Variable	Scope 1 + 2 Location Log	Scope 1 + 2 Market Log
Estimator	PanelOLS	PanelOLS
No. Observations	492	492
Cov. Est.	Clustered	Clustered
R-squared	0.2789	0.2706
R-Squared (Within)	-2.3998	-1.9782
R-Squared (Between)	0.2945	0.2895
R-Squared (Overall)	0.2755	0.2653
F-statistic	26.578	25.490
P-value (F-stat)	0.0000	0.0000
========================	==========================	=========================
const	14.272***	14.004***
	(21.930)	(18.980)
Asset Turnover	-1.0424	**-1.8324***
	(-1.0902)	(-1.7068)
Current Ratio	-0.0977	-0.1674
	(-0.1639)	(-0.3048)
EBITDA Margin, Percent	-0.0046	-0.0021
	(-0.2495)	(-0.0934)
ESG Score	3.133e-12	1.135e-10
	(0.0280)	(0.9219)
Quick Ratio	-0.6679	-0.6254
	(-1.0330)	(-1.0074)
Expense Ratio	0.3193	1.2150
	(0.3142)	(1.1034)
Utilities	**2.9078*** **	**3.0037*** **
	(4.3068)	(4.0634)
==========================	============================	===========================
Effects	Time	Time

Asset Turnover was only statistically significant for the third model at the 10% significance level. The negative coefficient is characteristically high meaning a 1 unit increase in Asset Turnover has an expected 183.24% decrease in Scope 1 + 2 Market Emissions. Utilities was statistically significant at the 1% significance level across both models with high positive coefficients.

Asset Turnover was statistically significant at the 10% significance level for a total of 3 models with highly negative coefficients. This only provides weak support for the hypothesis and one of the potential causes for this is “multicollinearity” as shown by the Variance Inflation Factor (VIF) values:

Table 10: Multicollinearity of Features

VIF	Features
12.9	const
29.0	Asset Turnover
4.7	Current Ratio
1.6	EBITDA Margin, Percent
2.6	ESG Score
4.6	Quick Ratio
30.6	Expense Ratio
1.2	Utilities

Asset Turnover is highly correlated with Expense Ratio which means it is less likely they will be statistically significant.

5 Conclusion

The relationship between Asset Turnover and firm Carbon Emissions, as described in the hypothesis, was studied in-depth. The results from the panel data regressions supported the existence of this relationship but not to a satisfactory level. Fortunately, these mildly disappointing results may not represent reality as there some key considerations that need to be made.

The models could benefit from the removal of Expense Ratio as this would allow the statistical significance of Asset Turnover to be accurately measured by removing multicollinearity. Another important point is the statistical significance of the Utilities dummy variable - it seems that the Sector of a company plays a large role in explaining carbon emissions.

The entity effects rendered the models worthless as the explanatory variables were absorbed and the time-effects of the PanelOLS models were jointly statistically insignificant at the 1% significance level:

Table 11: F-Statistic Test of Poolability

Model	**F-stat**	**P-value**
Model 1	0.3829	0.7654
Model 2	0.9307	0.4256
Model 3	3.0823	0.0271
Model 4	0.8644	0.4594
Model 5	1.4569	0.2255

This indicates that Panel OLS is not suitable for this data and PooledOLS could perform better. While there is a degree of support for the hypothesis to be derived from this analysis, the analysis is marred by issues which would need to be resolved before the hypothesised relationship can be clearly established (see Appendix A.1 for discussion of some solutions).

6 Bibliography

- Emmanuel Asafo-Adjei, A. M. A. L. A. B. A. S. &. R. M. G., 2022. Similarities among equities returns in multi-frequencies: insights from sustainable responsible investing. *Sustainable Finance & Investment,* Issue 12.

- Dan, 2023. Asset Structure, Asset Utilization Efficiency, and Carbon Emission Performance: Evidence from Panel Data of China"s Low-Carbon Industry. *Sustainability, Section: Economic and Business Aspects of Sustainability,* Volume 15(7), 6264; https://doi.org/10.3390/su15076264.

- Qian Xu, Y. L. H. L. B. L., 2021. Does corporate environmental responsibility (CER) affect corporate financial performance? Evidence from the global public construction firms. *Cleaner Production,* 315(128131).

- Houqe, O. H. A., 2022. The Effects of Carbon Emissions and Agency Costs on Firm Performance. *Risk Financial Management,* 15(4), 152.

- Bukit & Ginting, e. a., 2017. *Environmental performance, profitability, asset utilization, debt monitoring and firm value.* Medan, Indonesia, International Conference on Agriculture, Environment, and Food Security 7–8 November 2017.

A Appendix A - Pooled OLS & Ridge Regression Models

This appendix will examine the results of 5 Pooled OLS Models (Pooled OLS is a special case of OLS applied to panel data where cross-sectional and time-effects are discounted) where the Expense Ratio has been removed to hopefully negate the existence of multicollinearity. Additionally, the predictive power of the resulting model will be inspected through the use of Ridge Regression which is a penalized linear regression.

A.1 Pooled OLS Models

As an example of the regression output this appendix will include the regression results summary of the first model which has log(Scope 1 Emissions) as its dependent variable. All other outputs will not be provided but will be analysed in this section (for reference please see provided Python code):

Table 12: Summary of Regression Results for Model 1.

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	13.3392	0.354	37.670	0.000	12.643	14.035
Asset Turnover	-0.7828	0.204	-3.838	0.000	-1.184	-0.382
Current Ratio	0.1531	0.308	0.497	0.620	-0.453	0.759
EBITDA Margin	-0.0103	0.008	-1.344	0.180	-0.025	0.005
ESG Score	4.061e-11	8.96e-11	0.453	0.651	-1.35e-10	2.17e-10
Quick Ratio	-1.1794	0.330	-3.575	0.000	-1.828	-0.531
Utilities	3.9984	0.417	9.592	0.000	3.179	4.817

Utilities remains almost identical to the Panel OLS and further proves that Sector categorisation may be a necessary step to take. The issue with creating dummy variables for each sector is that it creates extremely high levels of multicollinearity both with each other and with the constant. However, the impact of sector categorisation seems integral to the performance of these models - therefore an elegant solution would be to utilise the unsupervised machine learning algorithm "Principal Component Analysis" to reduce the dimensionality of the data and counteract the multicollinearity.

Asset Turnover has now become statistically significant at the 5% signficance level for *all 5 models*. In addition to this, it is statistically significant at the 1% significance level for 4 of the models. The removal of Expense Ratio and change of technique to Pooled OLS seems to have rectified many of the issues that were faced initially as is shown by the new VIF values:

VIF	**Features**
11.80	const
1.30	Asset Turnover
4.60	Current Ratio
1.20	EBITDA Margin
1.00	ESG Score
4.60	Quick Ratio
1.10	Utilities

This table shows that multicollinearity has been mostly resolved as the VIF values are all below 5 (apart from the constant which is irrelevant since it is not a predictor). White's Test was also run on each of these models to check the heteroscedasticity of the residuals. For **all 5 of the models** the **residuals** were found to be **homoscedastic**.

While the modulus of the negative coefficients decreased massively for Asset Turnover, the hypothesised relationship has extremely strong support as it is now statistically significant in every model. The Pooled OLS Models themselves performed very similar to the Panel OLS models with the R-squared values being slightly higher - however the insight into the significance of each variable is much clearer.

A.2 Ridge Regression Models

Despite multicollinearity seemingly being solved, a good way to counteract it even further is to introduce the penalized linear regression "Ridge Regression". This regression technique will be used to further eliminate any issues of multicollinearity. The predictive performance of this model will also be assessed since a train-test-split was performed on the data prior to model creation.

5 Ridge Regression models were created in the same way as the Pooled OLS Model except the data was split 70:30 into training and test data. The dependent and independent variables remain the same. In order to assess the model after penalization, Mean-Squared Error and R-squared are the statistics that will be provided:

Model	MSE (Train)	MSE (Test)	R2 (Train)	R2 (Test)
Model 1	5.016	5.855	0.315	0.100
Model 2	2.717	2.705	0.221	0.141
Model 3	3.832	4.609	0.205	-0.014
Model 4	2.996	3.330	0.311	0.157
Model 5	3.519	3.980	0.308	0.107

The predictive performances of the Ridge Regression models are underwhelming to say the least. There is a huge disparity between MSE and R-squared from the training data to the test data (apart from Model 2). In fact, Model 3 even has a negative R-squared on the test data which suggests it is performing worse than the mean of the target values.

This likely means that the model is overfitting the training data and is not generalising well to new, unseen data (the test data). A potential cause is the fact that the number of observations far outweighs the number of features (predictor variables) in these models which means that the penalization is much less effective. One solution would be to introduce many more variables while keeping the same sample size. Another solution would be to forego use of the Ridge Regression and instead use a machine learning algorithms such as Random Forest or XGBoost as the aggregation method employed by these algorithms works to counteract overfitting.

While the model itself is not perfect, likely due to a lack of data, the hypothesised relationship is clearly established once certain measures are taken to address the issues in the Panel OLS Regressions and data.

Return to Table of Contents