User Guide
Greenhouse Gas Emissions Estimator

February 2024

Version E02/24

1. Background and Objectives

The Hong Kong University of Science and Technology cooperated with the Green and Sustainable Finance Cross-Agency Steering Group to build two greenhouse gases (GHG) emissions tools in bridging the data gap highlighted by the industry.

The GHG emissions estimators aims to facilitate financial institutions to estimate the Scopes 1 and 2 GHG emissions of a company based on the widely adopted international standards, such as Global GHG Accounting & Reporting Standard for Financed Emissions of the Partnership for Carbon Accounting Financials.

2. Introduction to the GHG Emissions Estimator

While GHG emissions reporting has improved in recent years, reported emissions data are still limited. The GHG emissions estimator aims to provide financial institutions with an alternative means to assess the GHG emissions of companies in their investment or loan portfolios when reported data are not available.

Our focus is on estimating the Scopes 1 and 2 GHG emissions with company data. These categories encompass the direct and indirect emissions resulting from a company's activities, including the combustion of fossil fuels and the generation of purchased electricity, heat, or steam.

GHG Emissions Estimation Model

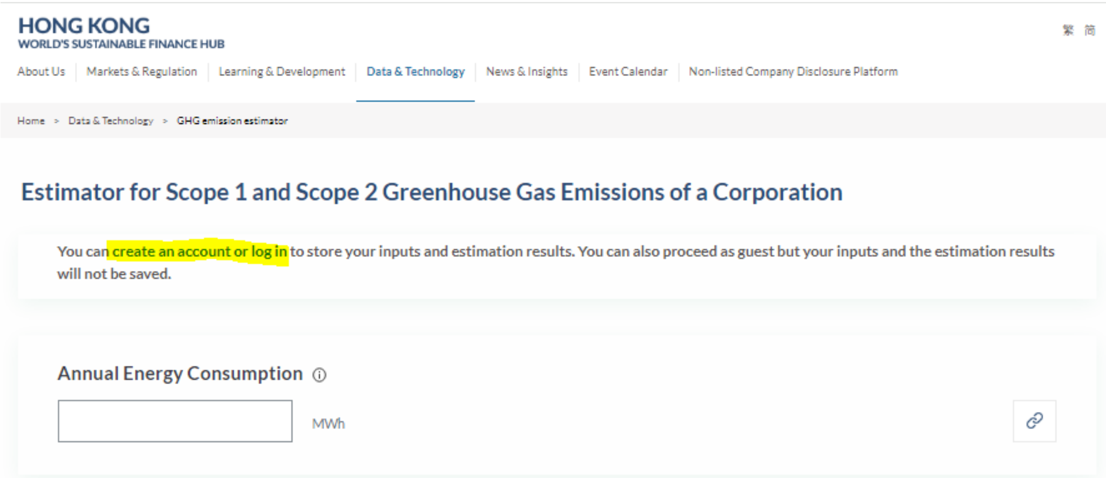
We have developed a regression model to estimate Scopes 1 and 2 GHG emissions based on energy consumption of companies. Using data of listed companies and small and medium-sized enterprises (SMEs), our objective is to leverage the power of regression to provide accurate and reliable estimates of Scopes 1 and 2 GHG emissions. By combining these datasets, we aim to capture a comprehensive representation of the corporate landscape in terms of energy consumption and associated Scopes 1 and 2 GHG emissions.

For the details of the GHG emissions estimation model, including the data collection, exploratory analysis, model testing and results, please read *section 4* ***"Technical Documentation of the GHG Emissions Estimator"***.

3. How to use the GHG Emissions Estimator

The GHG emissions estimator is a simple online tool which estimates GHG emissions based on a single factor, which is the energy consumption. You can refer to the icons in the tool for simple guidance on the input field. Please read below for details on using the GHG emission estimator.

You can create an account or log in to store your inputs and estimation results, please refer to section 3.3 for the functions of the account platform. You can also proceed as guest but your inputs and the estimation results will not be saved.

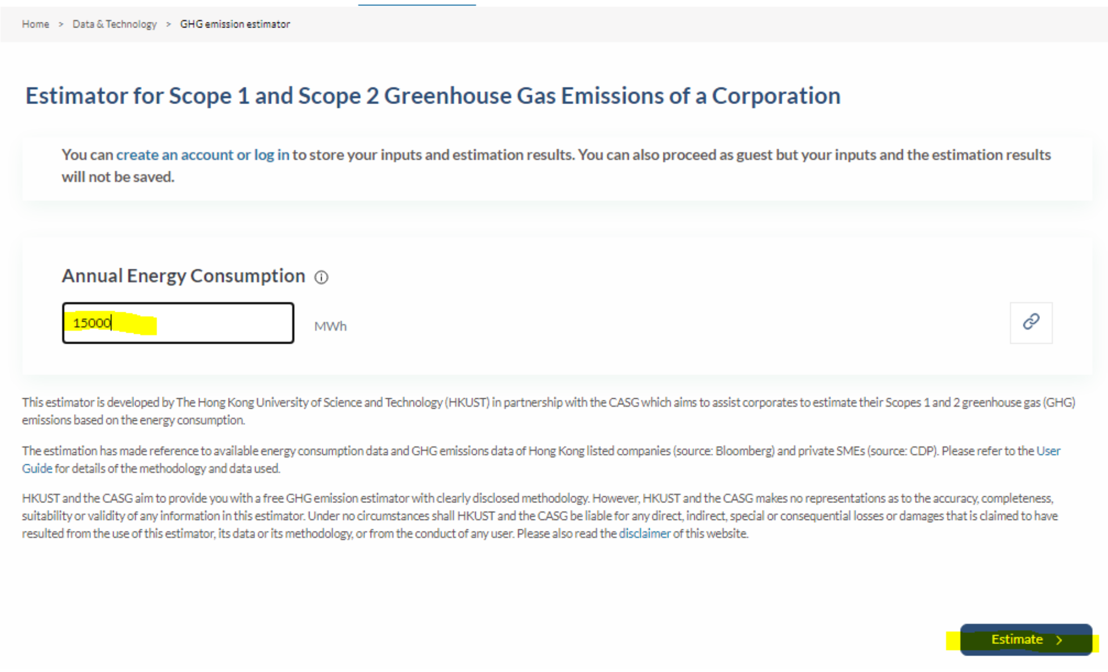


3.1 Inputting energy consumption

You can input the total energy consumption (in megawatt hour (MWh)), including fuel and electricity usage, during the reporting year. Please note that the model is sensitive to outliers, the energy consumption inputted should lie within the range of data covered by the model, which spans from 0.15 MWh to 520,000,000 MWh.

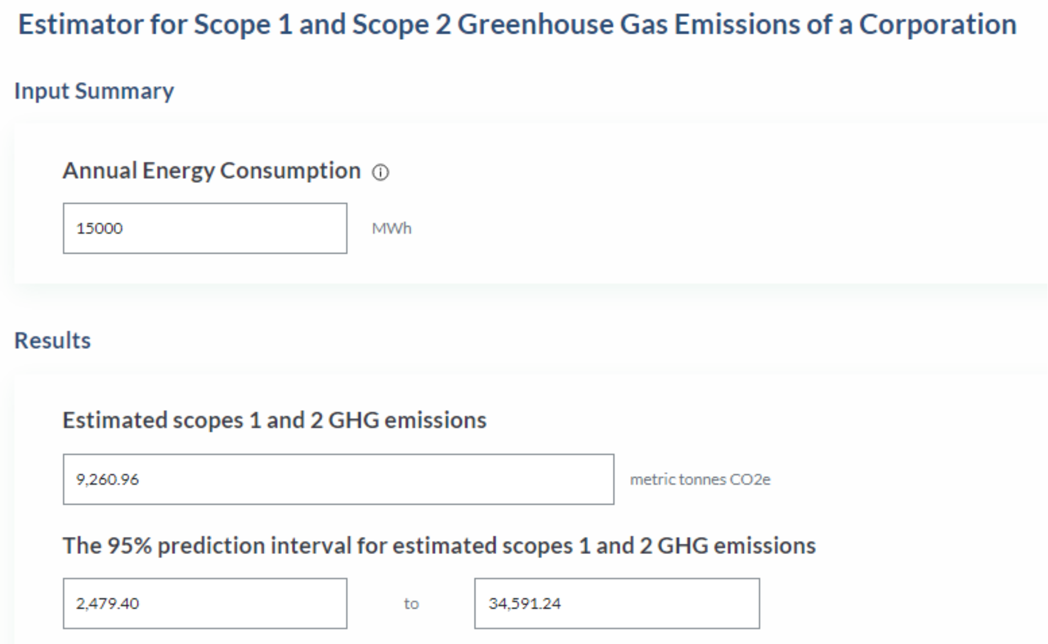
You may also refer to icon for a unit conversion table which may be helpful for performing unit conversion.

Then you can press "Estimate" to proceed to the estimation results.



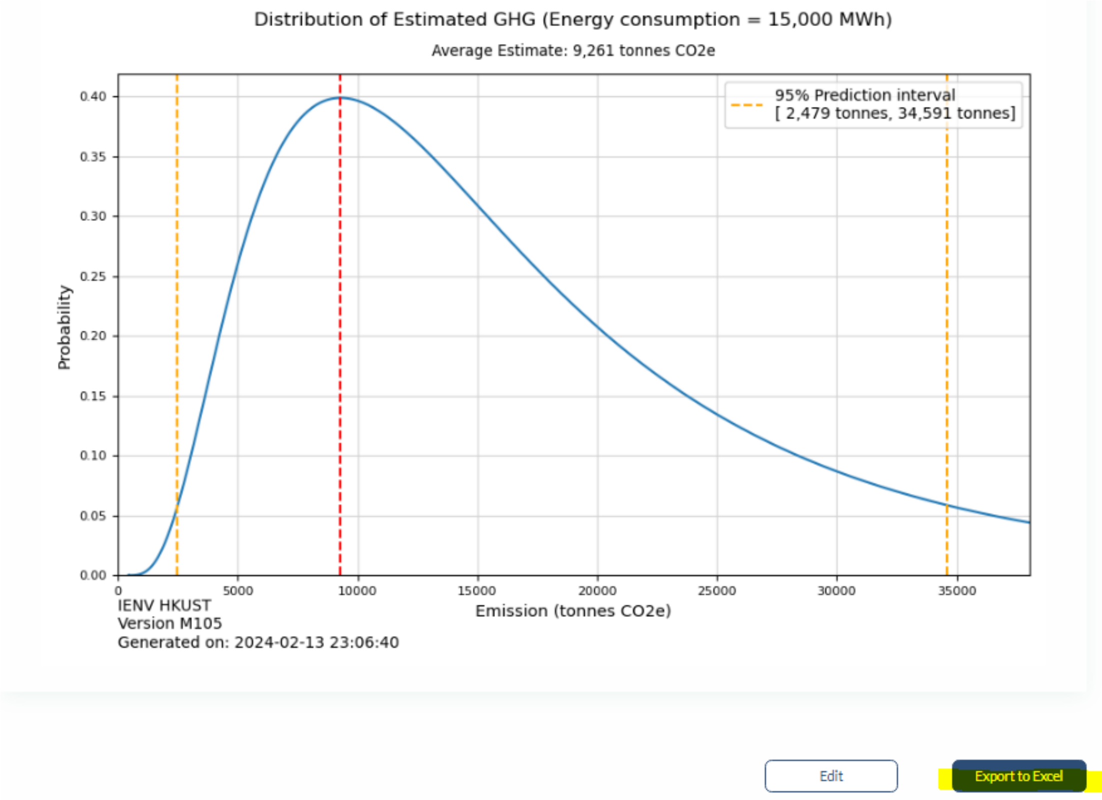
3.2 Reviewing results

On the results page, you can see the energy consumption input, the estimation results, including the 95% prediction interval for estimated Scopes 1 and 2 GHG emissions, and a plot of distribution of estimated Scopes 1 and 2 GHG emissions.



Prediction interval represents the uncertainty of predicting the value of a future observation from a population based on the distribution of a number of previous observations. For example, if the 95% prediction interval for estimated Scopes 1 and 2 GHG emissions at an annual energy consumption of 15,000 MWh is

[2,479.4, 34,591.24], you can be 95% confident that the Scopes 1 and 2 GHG emissions at an annual energy consumption of 15,000 MWh is between 2,479.4 and 34,591.24 tonnes CO2e.

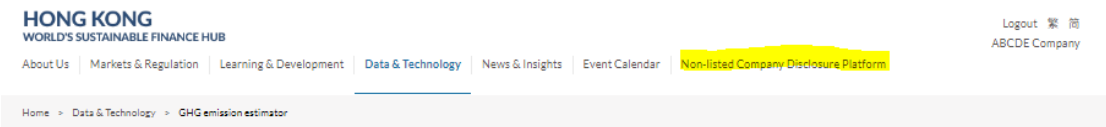


You can click the "Export to Excel" to download the energy consumption inputted and estimation results for record keeping.

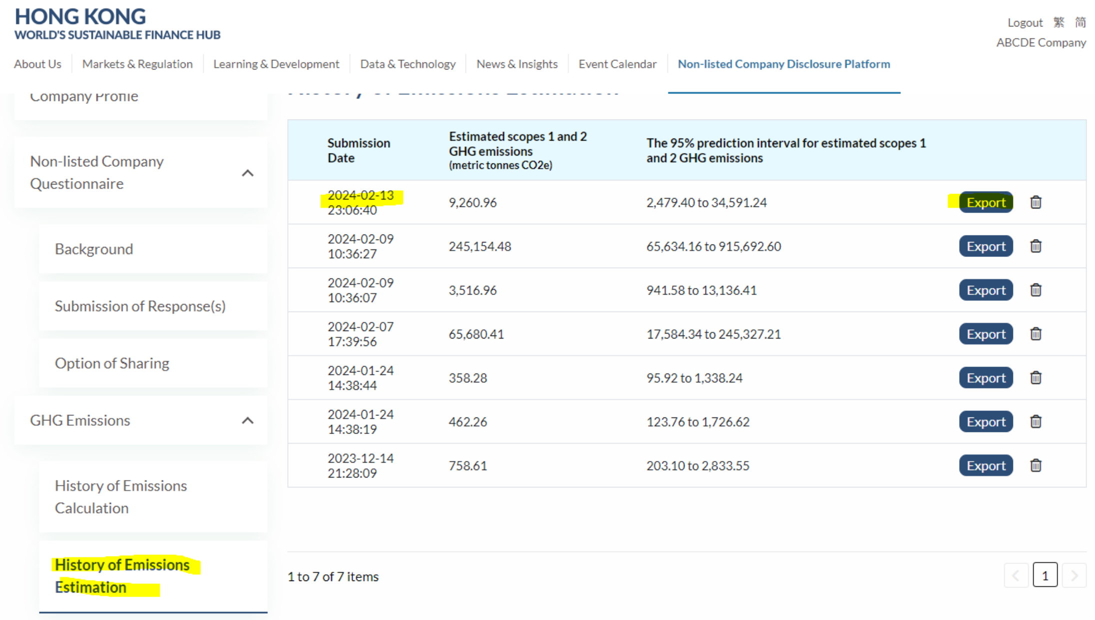
3.5 Retrieving estimation results (only for users who created an account)

If you create an account and log in, your inputs and estimation results are automatically saved in the account platform.

To access your user profile and retrieve your estimation history, press "Non-listed Company Disclosure Platform".

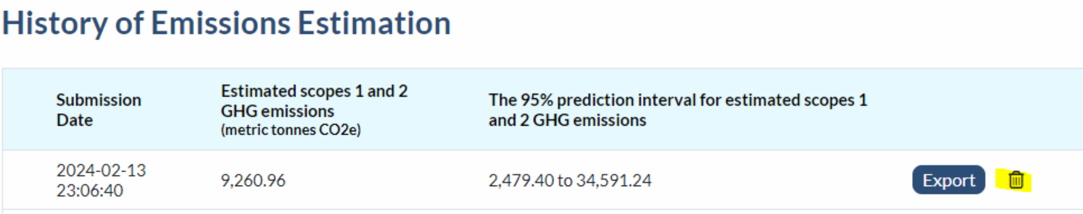


Then click "History of Emissions Estimation" on the menu, your estimation history will be shown. You can review the information online or click "Export" to download the respective estimation results for record keeping.



Submission Date	Estimated scopes 1 and 2 GHG emissions (metric tonnes CO2e)	The 95% prediction interval for estimated scopes 1 and 2 GHG emissions	
2024-02-13 23:06:40	9,260.96	2,479.40 to 34,591.24	Export
2024-02-09 10:36:27	245,154.48	65,634.16 to 915,692.60	Export
2024-02-09 10:36:07	3,516.96	941.58 to 13,136.41	Export
2024-02-07 17:39:56	65,680.41	17,584.34 to 245,327.21	Export
2024-01-24 14:38:44	358.28	95.92 to 1,338.24	Export
2024-01-24 14:38:19	462.26	123.76 to 1,726.62	Export
2023-12-14 21:28:09	758.61	203.10 to 2,833.55	Export

To remove an estimation history, click the  icon and confirm to delete.



Submission Date	Estimated scopes 1 and 2 GHG emissions (metric tonnes CO2e)	The 95% prediction interval for estimated scopes 1 and 2 GHG emissions	
2024-02-13 23:06:40	9,260.96	2,479.40 to 34,591.24	Export

4. Technical Documentation of the GHG Emissions Estimator

4.1 Data Collection and Preparation

The data used for this study were sourced from the Bloomberg database, specifically targeting Hong Kong listed companies from the years 2020 and 2021. The initial dataset consisted of 2,683 companies, each containing information such as sector classification, revenue, assets, operating expenses, EBITDA (earnings before interest, taxes, depreciation, and amortization), employee numbers, EVIC (enterprise value including cash), energy consumption, and Scope 1 and Scope 2 GHG emissions.

After the data collection phase, data cleaning procedures were performed to ensure data quality and relevance. As a result, 320 listed companies from the year 2020 and 359 listed companies from the year 2021, all based in Hong Kong, remained in the dataset for analysis.

To supplement the dataset, SME data was obtained from the CDP database, specifically focusing on publicly available responses for the year 2021. Out of the publicly available responses, 23 were private companies that provided complete disclosure on Scopes 1 and 2 emissions as well as energy consumption. To ensure the relevance of the data for the purpose of GHG emissions estimation for Hong Kong SMEs, the dataset was further refined based on the criterion of having at least part of their business operations in Hong Kong and a reported revenue of USD 100 million or less. Consequently, 9 companies met these criteria and were included in the final dataset.

In order to prepare for the subsequent regression analysis, the Scope 1 and 2 GHG emissions were aggregated to obtain the total GHG emissions, which would be referred to throughout this document. A log transformation was applied to all the values and they were validated to meet the assumptions of normal distribution required for the regression analysis.

4.2 Exploratory Analysis

To explore the relationship between the variables and GHG emissions, an exploratory analysis was conducted to the listed company data of 2021. Initially, the following variables were considered: sales and marketing, research and development, asset turnover, operating expenses, other operating expenses, general and administrative, EVIC, total asset, number of employees, Revenue, EBITDA, and Energy.

Variable Selection

In the variable selection process, the Pearson correlation coefficients were computed to analyze the relationship between each variable and the total GHG emissions. The correlation coefficient table is as follows.

Parameter (Logged values)	**Correlation coefficient**
Sales and marketing	0.11
Research and development	0.26
Asset turnover	0.27
Operating expenses	0.48
Other operating expenses	0.48
General and administrative	0.58
EVIC	0.64
Total asset	0.64
Number of employees	0.66
Revenue	0.71
EBITDA	0.75
Scope 1 GHG emission	0.93
Scope 2 GHG emission	0.96
Energy consumption	0.97
Total GHG emission	1.00

Table 1 Pearson Correlation Coefficients of Variables with Total GHG Emissions

Based on these correlation coefficients, the variables with the highest correlations to the total GHG emissions were selected for further analysis. The top six variables with the strongest correlations were identified as EVIC, Total asset, Number of employees, Revenue, EBITDA, and Energy consumption. These variables demonstrate a considerable positive relationship with total GHG emissions, with Energy consumption exhibiting the highest correlation coefficient of 0.97. By focusing on these selected variables, we can effectively estimate GHG emissions using financial metrics in the subsequent analysis.

Cluster Analysis

To overcome the limitation of having a relatively small number of Hong Kong listed companies for each individual sector, a cluster analysis was conducted to group similar sectors together. The analysis aimed to identify sectors that exhibited similar patterns in GHG emissions. Initially, there were 11 sectors in the dataset:

1. Communication Services
2. Consumer Discretionary
3. Consumer Staples
4. Energy
5. Financials
6. Health Care
7. Industrials
8. Information Technology
9. Materials
10. Real Estate
11. Utilities

The cluster analysis was performed by considering the similarity of mean, median, and standard deviation of GHG emissions within each sector. This approach allowed for the identification of sectors with comparable emission profiles. As a result, the 11 sectors were grouped into five major sector groups (Figure 1):

1. Financials
2. Communication Services, Real Estate, Consumer Discretionary, Health Care
3. Materials
4. Energy, Utilities
5. Industrials, Information Technology, Consumer Staple

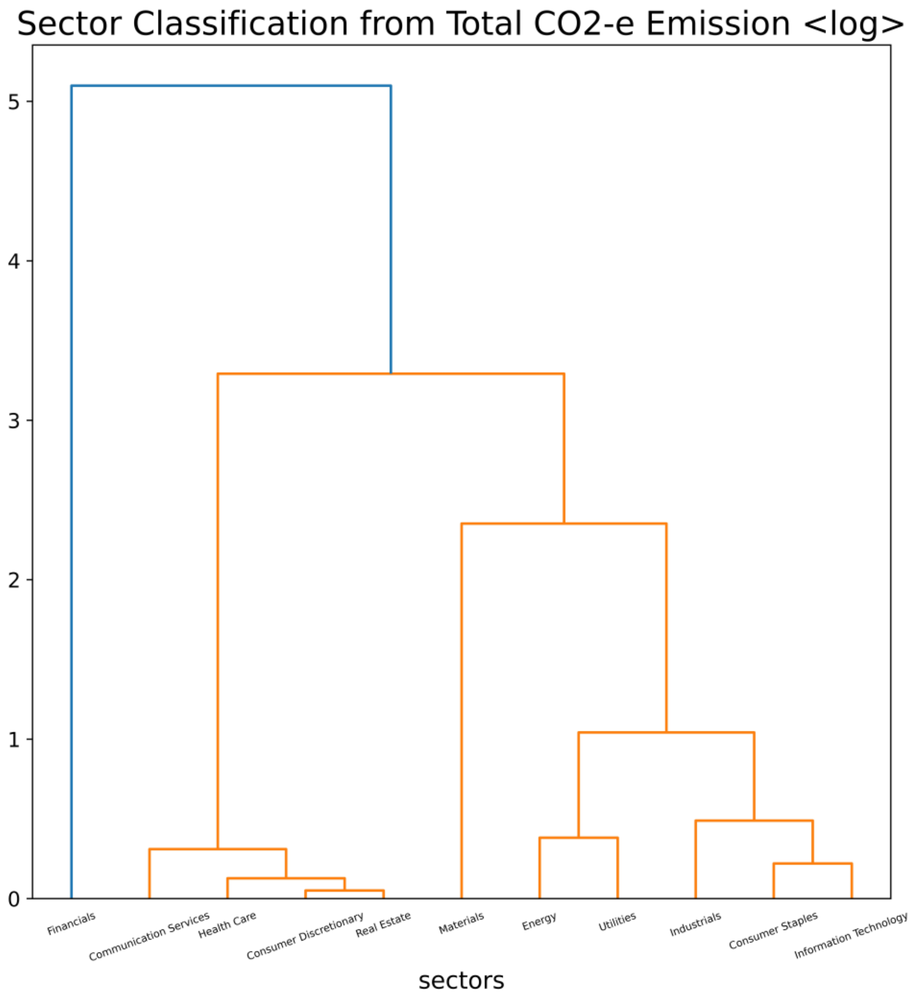


Figure 1. Dendrogram of illustrating the Results of Cluster Analysis for Sector Groups

By performing the cluster analysis, similar sectors were grouped together, allowing for a more robust regression analysis of GHG emissions using a larger combined dataset for each sector group. This approach enabled the exploration of sector-specific relationships with GHG emissions and facilitated more accurate predictions within each sector group.

Interrelationship of Variables

To examine the relationships between different parameters across different sectors, a pair plot analysis was conducted. The pair plot provided a visual representation of the interrelationship between the selected variables and their association with total GHG emissions. In addition, the pair plot helped identify any prominent factors influencing GHG emissions across sectors.

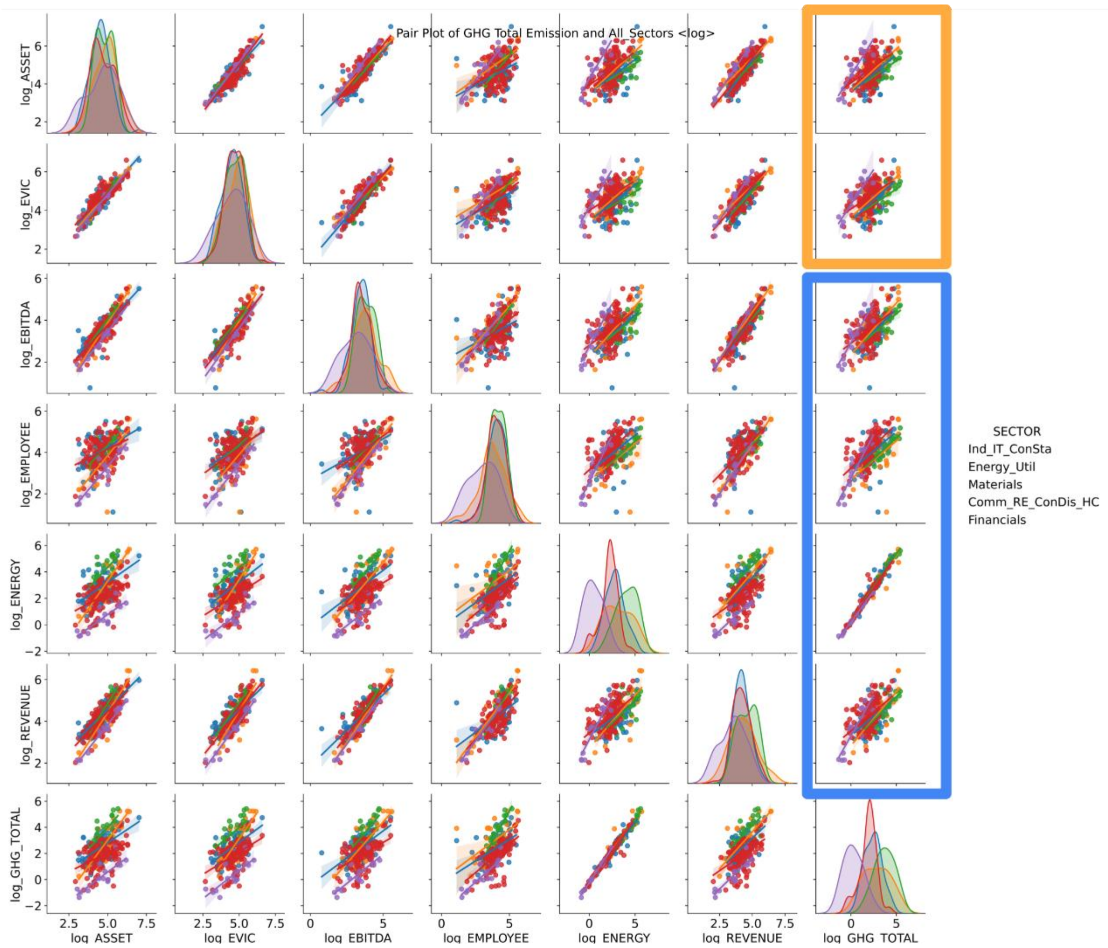


Figure 2. Pair plot of Logged GHG Emissions (tonnes) with Logged Energy Consumption (MWh)

In Figure 2, the pair plot displays the scatter plots and correlations between the variables. The rightmost column of the pair plot specifically focuses on the factors affecting total GHG emissions.

One notable observation from the pair plot analysis was the strong positive relationship between the variable "Energy" and total GHG emissions. This relationship was consistently observed across all sectors, indicating that energy consumption has a significant impact on GHG emissions regardless of the sector. The scatter plots in this column consistently showed an upward trend, reinforcing the positive correlation.

Additionally, it was observed that variables such as "Asset" and "EVIC" exhibited weaker correlations with GHG emissions compared to the other four parameters. The scatter plots for these variables appeared to be more sparsely spaced, indicating a less pronounced relationship with GHG emissions. As a result, "Asset" and "EVIC" were not considered in the subsequent regression analysis.

4.3 The Ordinary Least Squares (OLS) Regression Model

The choice of OLS regression as the modeling technique for estimating GHG emissions was made based on its suitability for the project's objectives and dataset. OLS regression is a commonly used statistical method that enables us to understand the relationship between a response variable and a set of predictor variables.

There are several advantages to using the OLS approach in this context. Firstly, OLS regression allows us to incorporate prior knowledge by selecting relevant financial metrics as predictor variables. This enables us to utilize existing knowledge about the impact of these metrics on GHG emissions. Secondly, OLS regression accounts for uncertainties by providing statistical inference, allowing us to assess the significance of the predictor variables and draw reliable conclusions (Whatley, 2022). Lastly, OLS regression provides prediction intervals, which quantify the uncertainty associated with the estimated GHG emissions, thereby providing a measure of the model's reliability.

Training and Testing the Model

All available listed company data from 2021 was used to train the regression model instead of splitting the data into training and validation sets. This approach was possible because the reported emissions could be directly verified against the estimated values. The direct verification of reported emissions added credibility to the training process and facilitated a thorough evaluation of the model's performance.

To assess the model's performance, various measures were employed, such as mean squared error and coefficient of determination. These metrics allowed for an evaluation of the model's ability to accurately estimate GHG emissions based on energy consumption.

Results

The OLS regression model was formulated to estimate GHG emissions based on the 2021 data of Hong Kong listed companies. The model included four predictor variables: energy, EBITDA, revenue, and employee. The response variable was the logarithm of GHG emissions.

The regression results for each sector group are as follows:

1. Financials:
   Log_GHG_Total = -0.067 + 1.105 * log_ENERGY + 0.006 * log_EBITDA + (-0.089) * log_REVENUE + 0.033 * log_EMPLOYEE

2. Materials:
   Log_GHG_Total = -0.953 + 0.873 * log_ENERGY + 0.048 * log_EBITDA + 0.046 * log_REVENUE + 0.170 * log_EMPLOYEE

3. Energy, Utilities:
   Log_GHG_Total = -0.337 + 0.898 * log_ENERGY + 0.086 * log_EBITDA + (-0.012) * log_REVENUE + 0.027 * log_EMPLOYEE

4. Industrials, Information Technology, Consumer Staples:
   Log_GHG_Total = -0.347 + 0.857 * log_ENERGY + 0.091 * log_EBITDA + 0.040 * log_REVENUE + (-0.027) * log_EMPLOYEE

5. Communication Services, Real Estate, Consumer Discretionary, Health Care:
   Log_GHG_Total = -0.359 + 0.944 * log_ENERGY + 0.012 * log_EBITDA + (-0.040) * log_REVENUE + 0.077 * log_EMPLOYEE

Based on the regression results, it can be concluded that energy remains the dominant factor when estimating GHG emissions for all sector groups, while the influence of other variables is negligible. Therefore, the decision was made to modify the regression model by considering only energy as the predictor variable and not differentiating sectors during the regression analysis.

The updated regression equation for estimating GHG emissions using the 2021 Hong Kong listed company data is as follows:

Log_GHG_TOTAL = 0.948 * log_ENERGY - 0.188

The R-square value of 0.962 indicates that approximately 96.2% of the variance in GHG emissions can be explained by the energy variable alone. This further supports the dominance of energy as a key factor in estimating GHG emissions within the dataset. By simplifying the model to focus solely on energy, we aim to improve the estimation accuracy while eliminating the need for sector divisions in the analysis.

4.4 Machine Learning Approach: Neural Network

As an alternative to the OLS regression analysis, a machine learning approach using a neural network model was considered to improve the accuracy of GHG emission estimation.

A neural network is a type of machine learning algorithm that uses multiple layers of interconnected nodes to model complex relationships between variables. It can handle non-linear relationships and capture interactions between variables, making it more effective for large and complex datasets. Whereas OLS regression is a statistical method that estimates the linear relationship between a dependent variable and one or more independent variables. It aims to minimize the sum of squared errors between the predicted and actual values of the dependent variable. The major advantage is that neural network model aimed to capture the implicit relationships between the input variables and GHG emissions, without assuming linearity. This allowed for the exploration of complex interactions and potential non-linear patterns that may exist within the data. By leveraging the power of neural networks, the model could potentially uncover hidden insights and improve the accuracy of GHG emission estimation.

The neural network model was designed with three layers and utilized eight input variables: energy, asset, employee, operating expenses, EVIC, EBITDA, ASSTO, and revenue. These variables were chosen based on their relevance to GHG emissions and availability in the dataset.

To train the neural network model, 80% of the entire 2021 dataset was allocated for training purposes. The remaining 20% would be used for testing the model's performance and evaluating its predictive capabilities. By splitting the data into training and testing subsets, it ensured that the model's performance could be assessed on unseen data, which is crucial for assessing its generalization ability.

To interpret the neural network model and understand the importance of different variables, Shapley Additive Explanation (SHAP) values were used. SHAP values provide a way to explain the model's output by attributing the contribution of each feature to the prediction. Figure 3 of the SHAP values analysis demonstrated that energy was the most influential factor among all the input parameters, again confirming its significance in estimating GHG emissions.

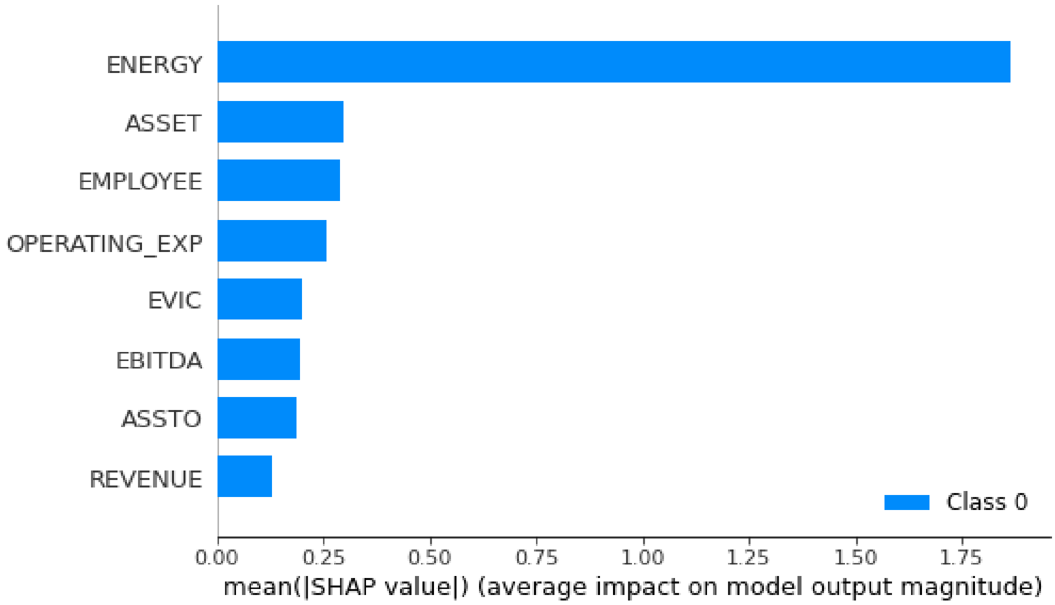


Figure 3. SHAP values of Parameters to the Total GHG Emissions

4.5 <u>Model Evaluation and Comparison</u>

To compare the performance of the OLS regression and neural network models, the root mean square (RMS) error was calculated. The RMS between the actual GHG emissions and the OLS regression results was 0.47, while the RMS between the actual GHG emissions and the neural network results was 0.75.

The lower RMS value of the OLS regression model indicates that its predictions are closer to the actual GHG emissions compared to the neural network model, indicating that the regression model actually performed better than the neural network model for the limited available data. As a result, it was decided to continue with the regression approach for GHG emission estimation in this project.

While the neural network model was not ultimately used in the analysis, it remains an intriguing avenue for future research. With a larger and more diverse dataset, the neural network model could potentially yield better results by capturing complex relationships and uncovering non-linear dynamics between the input variables and GHG emissions.

The subsequent sections of the document will delve into the incorporation of additional data and the impact of the expanded analysis on the accuracy of GHG emission estimation using the regression approach.

4.6 Application of the OLS Regression Model to SMEs

To assess the applicability of regression model to SMEs, additional data from CDP was incorporated. The focus was on SMEs operating in Hong Kong, with the criteria defined as companies having at least a portion of their business operations in Hong Kong and reporting a revenue of USD 100 million or less. Following this criteria, a total of 9 companies that met the specified conditions were identified from the year 2021.

An OLS regression analysis was performed on this subset of SMEs. Similar to the estimation results obtained for listed companies, the analysis revealed that energy remained a significant predictor variable in determining GHG emissions for SMEs. The coefficient associated with the energy predictor variable was calculated to be 0.949. Furthermore, the regression model demonstrated a strong fit, with an R-squared value of 0.979. This indicates that approximately 97.9% of the variance in GHG emissions can be explained by the energy-related factors considered in the analysis. These results reinforce the robust relationship between energy consumption and GHG emissions for both listed companies and SMEs in Hong Kong.

The regression equation for estimating GHG emissions using 2021 Hong Kong SME data:

Log_GHG_TOTAL = 0.949 * log_ENERGY - 0.064

Considering the promising results obtained from the OLS regression analysis on both listed companies and SMEs, the decision was made to combine the data sets. By grouping the Hong Kong listed company data and the SME data together, the aim was to leverage the increased sample size and diversity of companies to further enhance the accuracy of GHG emissions estimation.

The regression equation for estimating GHG emissions using the 2020 and 2021 Hong Kong listed company and 2021 SME data is as follows:

Log_GHG_TOTAL = 0.924 * log_ENERGY - 0.108

This decision was based on the understanding that the energy variable consistently emerged as a strong predictor of GHG emissions in both the listed companies and the SMEs. In the new regression analysis conducted using the combined data, the coefficient associated with the energy predictor variable was estimated to be 0.9493, indicating its significant influence. Additionally, the high R-squared value of 0.9789, which quantifies the proportion of variance in GHG emissions explained by the model, further supported the robustness of the regression model.

4.7 <u>Results and Analysis</u>

The analysis of GHG emissions estimation has yielded valuable insights into the emissions profiles of the companies under study. The estimated emissions for each company have been presented, highlighting the environmental impact of their operations. The regression model employed in this analysis has exhibited strong performance, with an impressive R-squared value of 0.949. This indicates that the model explains around 94.9% of the variance in the GHG emissions data, emphasizing a robust relationship between the predictor variables and the estimated emissions.

For example, a financial company in our data reported a total energy consumption of 47,370 MWh in 2021, resulting in actual Scopes 1 and 2 emissions of 31,870 tonnes. Through our regression model, we estimated an average GHG emissions value of 26,797 tonnes, representing a 16% difference from the actual emissions. Despite this deviation, it is noteworthy that the estimated emissions fall within the 95% prediction interval ranging from 7,174 tonnes to 100,090 tonnes. This indicates that our model effectively captures the uncertainty associated with the estimation process, providing a reasonable range within which the actual emissions are likely to lie.

The lineplot in Figure 4 illustrates the relationship between logged GHG emissions (tonnes) and logged energy consumption (MWh). The graph includes the 95% confidence interval of the coefficient, the 95% prediction interval, and the OLS result. The regression equation used to estimate the logged GHG emissions is y = 0.108 + 0.924x, with an R-squared value of 0.949, indicating a strong relationship between the logged energy consumption and the logged GHG emissions.

When examining the probability distribution of the logged estimated GHG emissions based on the logged energy consumption (Figure 5), the average estimate is $10^{4.428}$ tonnes, with a 95% prediction interval range spanning from $10^{3.856}$ to $10^{5.000}$ tonnes.

Figure 6 presents the probability distribution of the GHG emissions estimated using the energy consumption. The graph showcases the 95% prediction interval range, which varies from 7,174 to 100,090 tonnes. Moreover, the average estimate of the GHG emissions is 26,797 tonnes, representing the central tendency of the estimated emissions. This distribution provides further insights into the uncertainty and potential range of GHG emissions estimated based on energy consumption.

Furthermore, when comparing the average estimates from the regression model to the actual GHG emissions of all the companies studied, a root mean square value of 0.37 is obtained. This metric serves as a measure of the average difference between the estimated and actual emissions across the dataset. The relatively low RMS value suggests that the regression model provides reasonably accurate estimates, contributing to the overall reliability of the results.

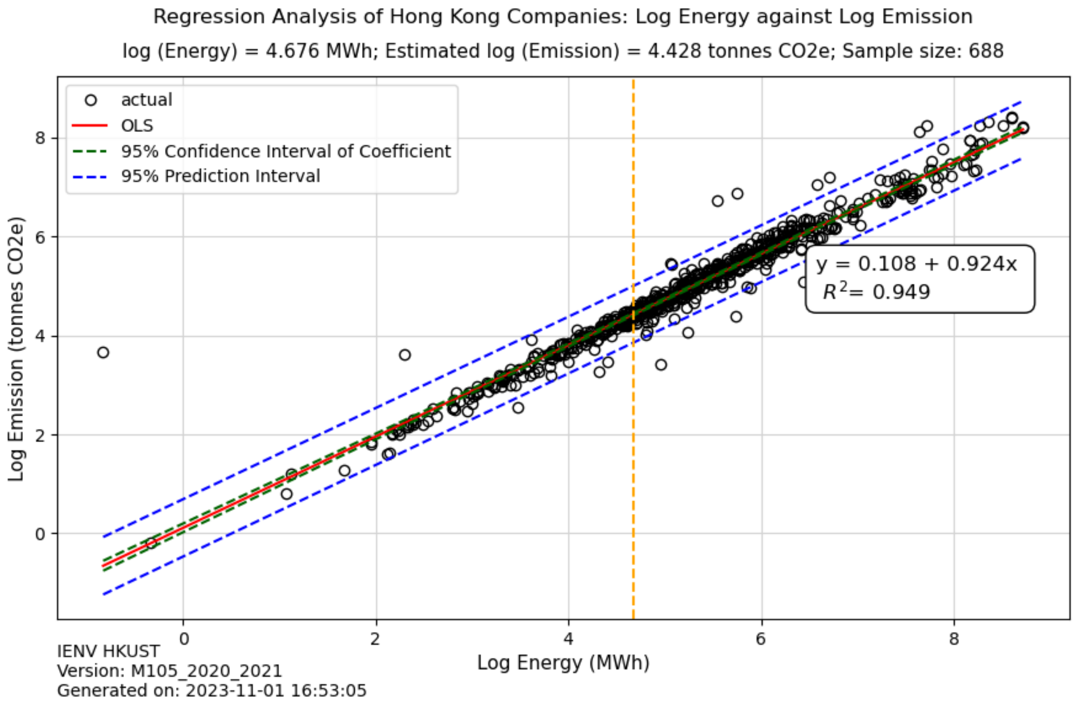


Figure 4. Lineplot of Logged GHG Emissions (tonnes) Estimated with Logged Energy Consumption (MWh)

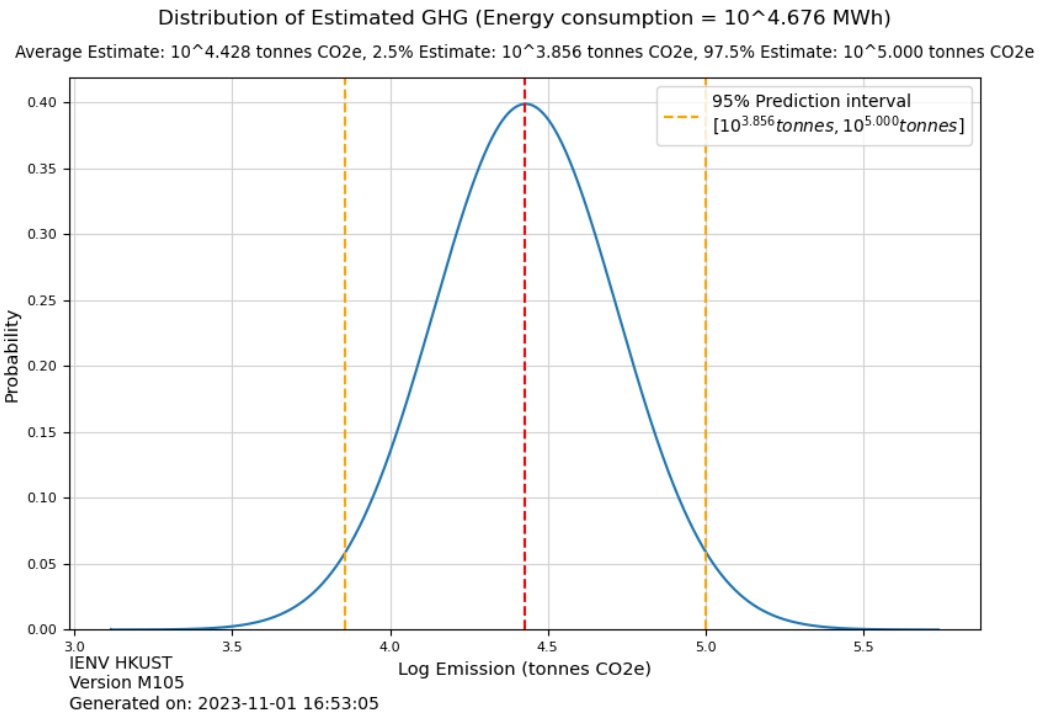


Figure 5. Probability Distribution of Logged GHG Emissions (tonnes) Estimated with Logged Energy Consumption (MWh)

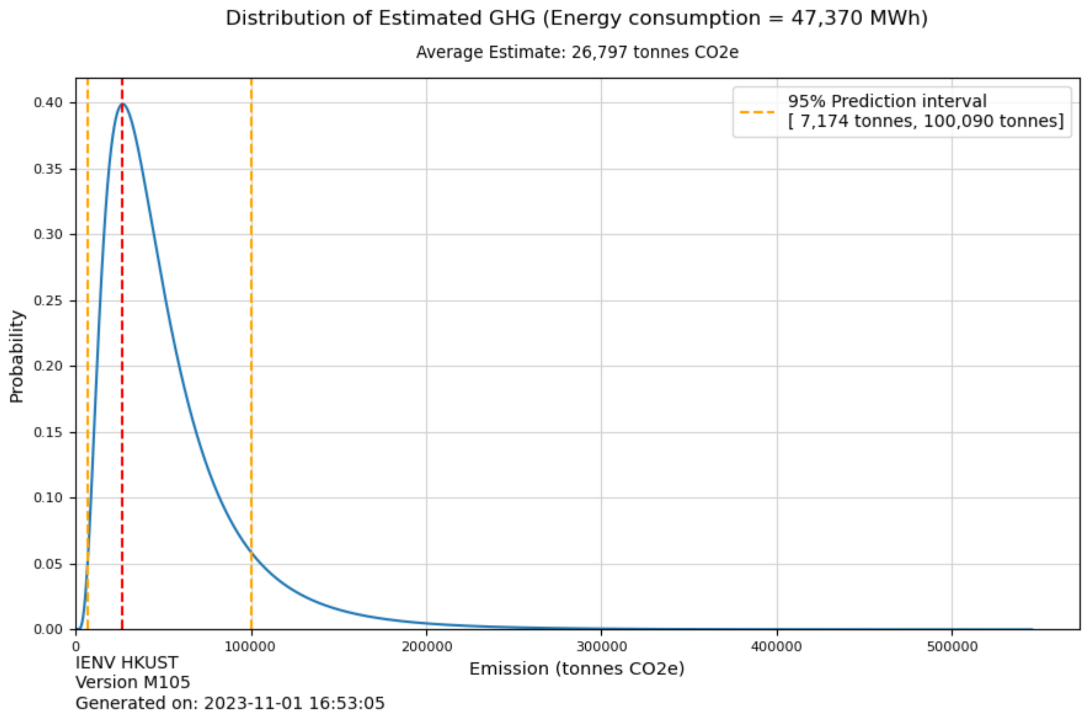


Figure 6. Probability Distribution of GHG Emissions (tonnes) Estimated with Energy Consumption (MWh)

In summary, the analysis of GHG emissions estimation has provided valuable insights into the emissions profiles of the studied companies. The regression model demonstrates a strong performance, with high explanatory power. The estimated emissions, although deviating from the actual values, fall within a reasonable range, capturing the uncertainty associated with the estimation process. These findings contribute to a more comprehensive understanding of the environmental impact of the companies under study and offer valuable information for future sustainability initiatives.

4.8 Limitations and Future Work

While our GHG emissions estimation model provides valuable insights, it is important to acknowledge certain limitations and consider areas for future improvement.

One limitation stems from the availability and quality of data used in the modeling process. It is essential to note that our dataset includes a limited number of companies. From the years 2020 and 2021, we had a total of 688 companies in our dataset, including both listed companies and SMEs. This limited sample size might introduce some limitations to the generalizability of the findings and the accuracy of the estimates. Expanding the dataset to include a larger and more diverse set of companies would enhance the accuracy and representativeness of our model. Moreover, the accuracy of our estimates relies heavily on the accuracy and completeness of the input data. Any biases or errors present in the data might impact the reliability of the results. In addition, our model assumes that the relationships between the predictor variables and GHG emissions remain consistent over time, which might not always hold true.

To enhance the model in the future, several avenues for improvement can be explored. Firstly, incorporating additional variables could provide a more comprehensive understanding of the factors influencing GHG emissions. For instance, including data on company-specific initiatives or operational practices could contribute to more accurate estimates. Furthermore, exploring alternative modeling techniques, such as machine learning algorithms or time series analysis, might unveil new insights and potentially improve the predictive accuracy of the model.

It is crucial to emphasize the ongoing need for data collection and model refinement. As new data becomes available, updating the model with the latest information can lead to more accurate estimations. Regularly evaluating and validating the model against actual emissions data can help identify any discrepancies or areas for improvement. Continuous monitoring and refinement of the model will contribute to its robustness and ensure that it remains relevant in an ever-changing business and environmental landscape.

4.9 Conclusion

In conclusion, this project has made significant findings and contributions by utilizing regression to estimate GHG emissions using energy consumption data. The key findings and contributions of this project highlight the effectiveness of the regression model in accurately estimating GHG emissions, with energy consumption demonstrating the highest correlation with total emissions. By leveraging this model, decision-makers can gain valuable insights into the environmental impact of energy consumption and make informed choices to promote sustainability.

The estimated GHG emissions have prominent implications for environmental sustainability and decision-making. Understanding the emissions associated with energy consumption enables businesses and policymakers to identify areas for improvement and develop targeted strategies to reduce their carbon footprint. This information can guide the adoption of energy-efficient practices, renewable energy sources, and emission reduction initiatives, ultimately contributing to mitigating climate change and promoting a more sustainable future.

This project underscores the importance of ongoing research and collaboration in the field of GHG emissions estimation. Continued efforts to refine models, incorporate additional variables, and explore alternative techniques will enhance the accuracy and applicability of estimation methods. Collaboration between researchers, industry stakeholders, and policymakers is essential to share knowledge, exchange best practices, and collectively work towards achieving global emission reduction targets.

4.10 Reference

Whatley, M. (2022). Ordinary Least Squares Regression. *In Introduction to Quantitative Analysis for International Educators* (pp. 91–112). Springer International Publishing AG. https://doi.org/10.1007/978-3-030-93831-4_7