Modeling Global Ecological and Socio-Economic Factors (2015-2023)

IDS702, Fall 2024

Meron Gedrago, Xiangyu Wang, Adil Gazder, Hongyi Duan

Abstract

This study explores the interconnected roles of education, income, and socioeconomic characteristics in shaping global carbon emissions. Using comprehensive global datasets, the research examines how variations in compulsory education duration impact carbon emissions across nations with differing income levels and population sizes. It also investigates the influence of life expectancy and labor force participation on a country's income group, highlighting indirect effects on environmental outcomes.The analysis seeks to uncover patterns and relationships that may provide insights into designing policies that simultaneously address climate change and socioeconomic disparities. By focusing on the interplay of education and income in mitigating carbon emissions, this study aims to contribute to a deeper understanding of sustainable development pathways and their implications for achieving global climate goals. Our work shows us that both life expectancy and labor force ratio are important factors in determining a country's income group status, affecting it to various extents from lower middle income to high income groups. We also saw that income and population size were influential predictors of carbon emissions, along with compulsory education duration also playing a critical role, where a country's income affected the carbon emission more if the country had lower compulsory education.

Introduction

Carbon emissions, primarily in the form of carbon dioxide (CO_2) released through the burning of fossil fuels, deforestation, and industrial activities, are the primary drivers of global warming. These emissions contribute to the accumulation of greenhouse gases in the atmosphere, trapping heat and causing a rise in global temperatures. The consequences of this warming include more frequent and severe weather events, rising sea levels, and widespread disruption to ecosystems and human livelihoods[1].

Global warming presents an urgent challenge as its effects are disproportionately felt by low-income nations, which often lack the resources to adapt to climate change. At the same time, high-income countries with historically higher carbon emissions bear significant responsibility for the crisis[2]. This imbalance underscores the necessity for policies that address both environmental sustainability and economic equity. Education plays a crucial role in this effort, equipping individuals with the knowledge and skills to develop innovative solutions, adopt sustainable practices, and influence policy decisions[3].

Despite growing awareness of the climate crisis, global carbon emissions have continued to rise in recent years[4]. This trajectory threatens progress toward international climate goals, such as those outlined in the Paris Agreement, and underscores the need for deeper exploration of the factors driving emissions and their interconnections with education and income. Addressing these issues requires understanding how education can mitigate carbon emissions and how income inequality influences both education outcomes and environmental impacts.

Building on these concerns, this study aims to investigate the interplay between education, carbon emissions, and income on a global scale. Specifically, it seeks to answer the following questions:

1. How do National income and population size influence carbon emissions across countries for different compulsory education durations ?
2. How do life expectancy and labor force (population able to work) affect the the income group of a country, which could indirectly affect carbon emission?

Through an analysis of global datasets encompassing socioeconomic and environmental indicators, this study aims to provide insights into the dynamic relationship between compulsory education, carbon emissions, and socioeconomic characteristics of citizens in these countries.

Methods

Data Sourcing and Pre-processing

Our data was a curated dataset[1] from the World Development Indicators (WDI) platform. WDI is the primary World Bank collection of development indicators, compiled from officially recognized international sources. It presents the most current and accurate global development data available, and includes national, regional and global estimates. The database contains 1,400+ indicators for more than 215 countries, with data for many indicators going back more than 50 years. We restricted our data to only our variables of interest and associated factors which may affect our research questions.

Variable Selection

The two main outcome variables we use is the total carbon emissions for a country for a given year and the Income group of a country for a given year[5]. The predictors for each of the research questions were decided based on prior variable research done and these are defined below:

- The Total CO_2 Emissions variable is measured in terms of Mega-tonnes of CO_2 equivalent (1 million metric tonnes) and is the total annual emissions of carbon dioxide equivalent units from human activities, excluding its use for Land Use, Land Use Change and Forestry (LULUCF). We only restrict the total CO_2 emissions to focus only on emissions from direct anthropocentric activities and exclude natural or semi-natural processes.
- Income Group is defined as the classification of a country into one of the of four categories (Low Income (L), Lower Middle Income (LM), Upper Middle Income (UM) and High Income (H)), based on its Gross Net Income per capita (measured in US$) for that given year. Based on dynamic thresholds which vary per year, each country is classified into one of the following four categories.

- National income measures non inflation adjusted and is defined as the difference between Gross National Income and the consumption of fixed capital and natural resources depletion.
- Compulsory Education Duration is the total number of years that children are legally obligated to attend school. For the purpose of this analysis, we have converted this variable into two. If the compulsory education years are below or equal to 10, it is classified as 'Low' and classified as 'High' otherwise.
- Labor force comprises people ages 15 and older who supply labor for the production of goods and services during a specified period. It includes people who are currently employed and people who are unemployed but seeking work as well as first-time job-seekers. Not everyone who works is included, however. Unpaid workers, family workers, and students are often omitted, and some countries do not count members of the armed forces.
- Labor force ratio is the overall ratio of the labor force to the total population of a country for a given year.

Model fitting and Evaluation

We utilized two distinct models to address our research questions: a multiple linear regression to analyze carbon emissions and a multinomial logistic regression to predict income group classifications.

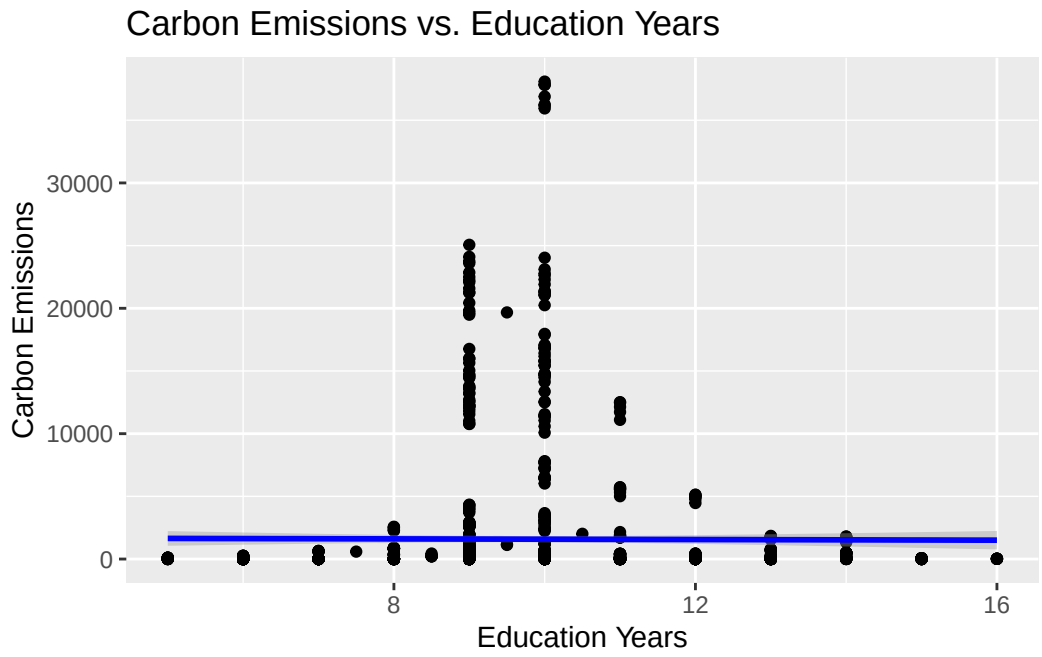
The multilinear regression model analyzing carbon emissions against `Education_new_fac`, net income, and population demonstrates strong predictive power, with an R^2 of 0.9329. The model's residual standard error is 1240, reflecting the average deviation of predictions from actual values. An ANOVA comparing models with and without interaction between `Education_new_fac` and `Net_Income` highlights a significant improvement in fit when interactions are included ($p < 0.0001$). This suggests that the effect of education on emissions is influenced by income levels, emphasizing the importance of considering predictor interdependencies.

The multinomial logistic regression model demonstrates a reasonable level of performance in predicting the income classification (L, LM, UM, H) with an overall accuracy of 62% and a 95% confidence interval of (59). This accuracy significantly exceeds the no-information rate (NIR) of 31, with a p-value of less than 2.2×10^{-16}, confirming that the model provides meaningful predictions beyond random chance. However, the model struggles more with classes LM and L, showing sensitivities of 42.19 and6.22 60.00, respectively. This indicates that the model often misclassifies countries in these groups, as reflected in the lower positive predictive values of 41.54 and 72.00. The balanced accuracy metrics highlight that the model is better at distinguishing between classes with higher representation (H, UM) while showing weaker performance for underrepresented groups.

Results

Research Question 1: Factors Affecting Carbon Emission

According to the exploration of the `Education Years` and `Carbon Emissions`, we could find out that the median of the `Education Years`(10 years) is quite a critical boundary when it comes to the relation to `Carbon Emissions`, so before we dig deeper for the corelationships, we will factor the `Education Years` into catagorical variables as `Education_new_fac`.



The data as shown below in 'Table 1' shows big differences between countries in carbon emissions, income, education, and population. A few countries produce a lot of carbon emissions, raising the global average, while most emit much less. Income is also uneven, with some very rich countries pulling the average far above what most others earn. Education levels are more even, with most countries requiring about 10 years of school, but improving the quality of education may help poorer nations. Bigger, richer countries tend to have larger populations and produce more carbon, but population size alone doesn't explain the differences.

Variables	Median..Q1..Q3.	Mean..SD.
Carbon Emission	39.39 [7.14, 447.70]	1584.21 (SD)
Compulsory Education	10 [9, 11]	9.78 (SD)
National Income (billion USD)	82.93 [13.22, 944.3]	2651 (SD)
Population (million)	17.34 [4.39, 126.7]	391.3 (SD)

	Carbon Emission	Compulsory Education	National Income (billion USD)	Population (million)
Median	39.39	10	82.93	17.34
1st Quartile	7.14	9	13.22	4.39
3rd Quartile	447.70	11	944.3	126.7
Mean	1584.21	9.78	2651	391.3

When we further explore the relationship of the four variables, we can see a strong positive association exists between Population and Carbon Emission, indicating that countries with larger populations tend to emit more carbon, likely due to increased industrial and energy demands. Similarly, National Income shows a positive trend with Carbon , suggesting that wealthier countries might contribute more to emissions. For countries with lower Compulsory Education, the relationship between National Income and Carbon Emission is stronger when compared to countries with higher Compulsory Education. Overall, all three variables look to be related to Carbon Emission from the scatterplot.

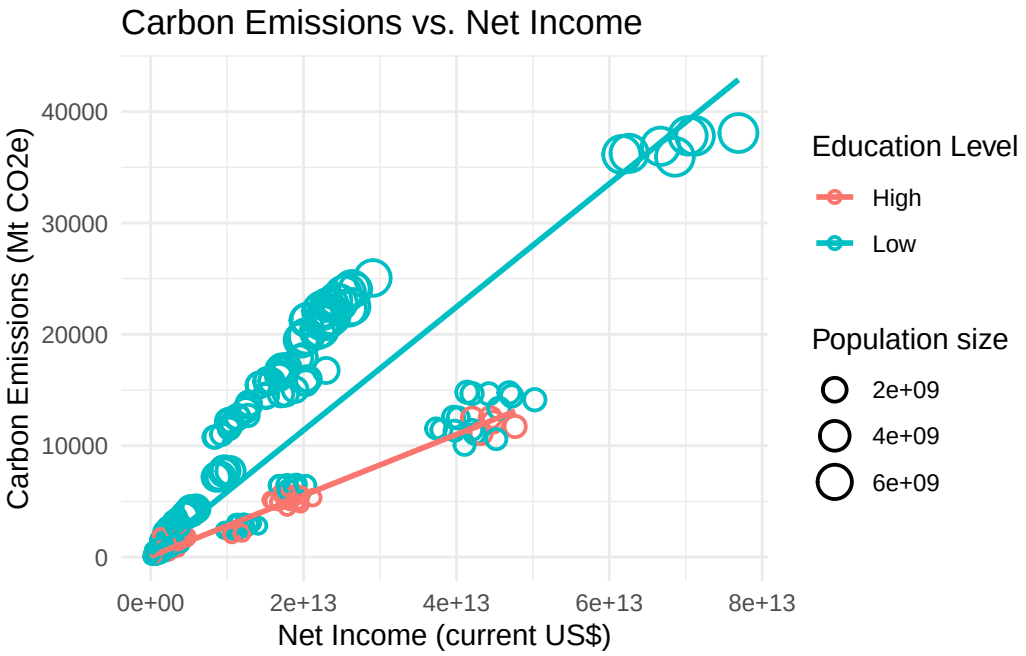
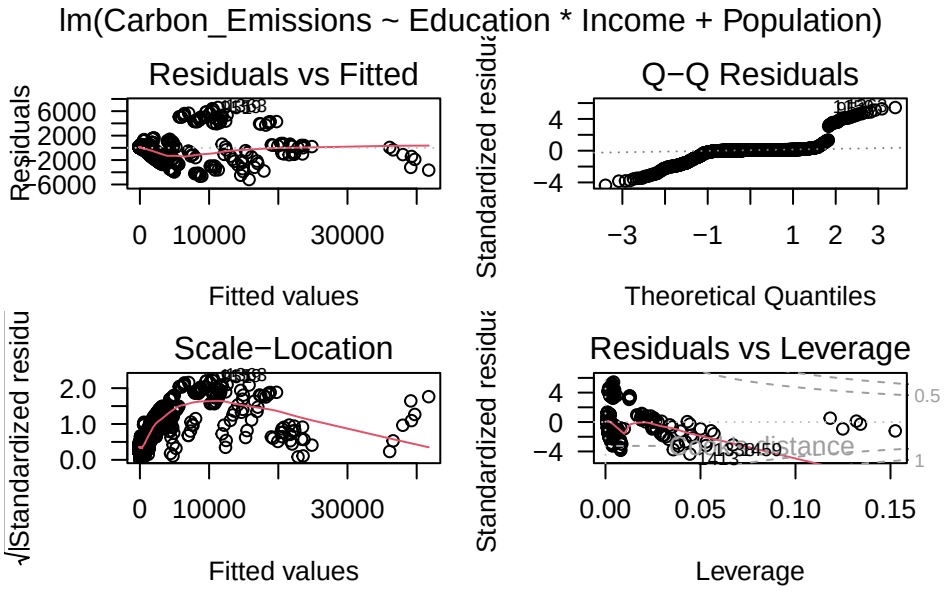


Table 2: Multiple Linear Regression Model Summary

Variable	Estimate	SE	t-value	p-value
Intercept	-34.04	61.44	-0.554	0.58
Compulsory Education	-110.7	61.44	-1.504	0.133
National Income	2.040e-10	1.034e-11	19.728	<0.001
Population	2.471e-6	4.666e-08	52.966	<0.001
Compulsory Education * National Income	8.714e-11	1.168e-11	7.463	<0.001

The regression model shows a strong overall fit, indicating that 93.52% of the variation in carbon emissions is explained by Compulsory Education, National Income and Population.The regression model shows how factors like education, income, and population affect carbon emissions. It tells us that net income and population size are very important in predicting carbon emissions—higher income and more people lead to more carbon emissions. From the results, countries with lower education seem to have a stronger link between income and carbon emissions. This means that while income and population are the biggest drivers, education can still influence how these factors interact with emissions.

We have plotted the diagnostic plots below to assess whether any assumptions of linear regression were violated and to identify influential points. From the Q-Q plot, the points are not perfectly aligned along the 45-degree line, but the deviation is within an acceptable range. In the residuals vs. fitted plot, there is a noticeable pattern in the residuals, though it does not form a specific shape; this may be due to country groups with similar characteristics that we have not accounted for in the model. Additionally, there are no influential points that exceed the established thresholds in the plots.

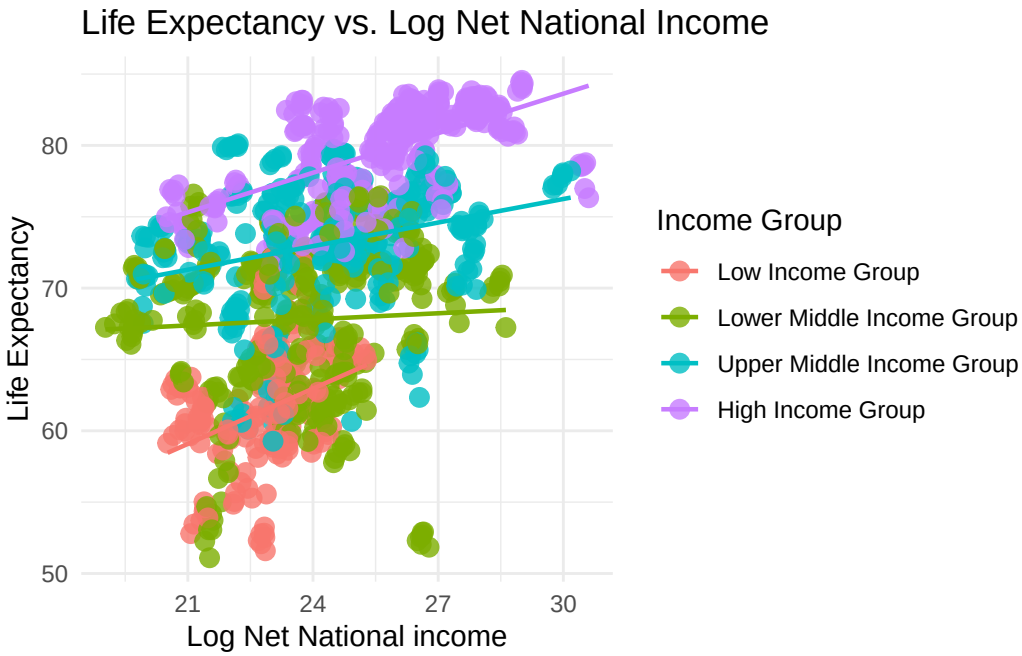


The factor model highlights that the relationship between education years and carbon emissions could have critical thresholds where education systems or societal behaviors influence emissions. For policy-making, focusing on specific education levels rather than increasing education years linearly might yield better outcomes for carbon reduction strategies.While both models perform well, the factor model provides deeper insights and better predictive power, making it more suitable for nuanced analyses and interventions.

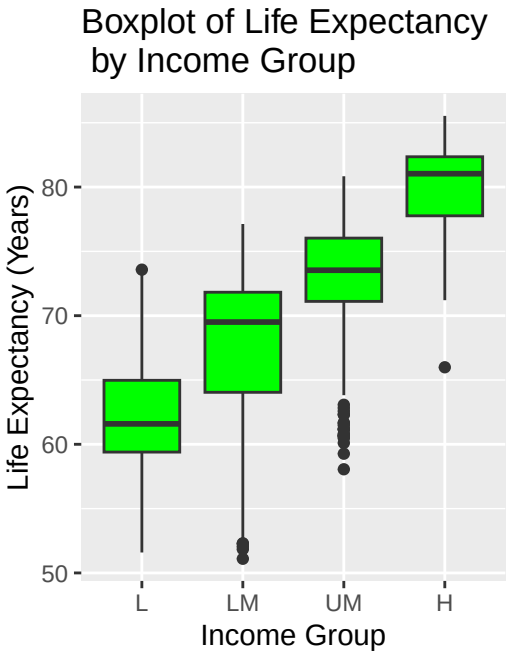
Research Question 2: Factors affecting Income Group

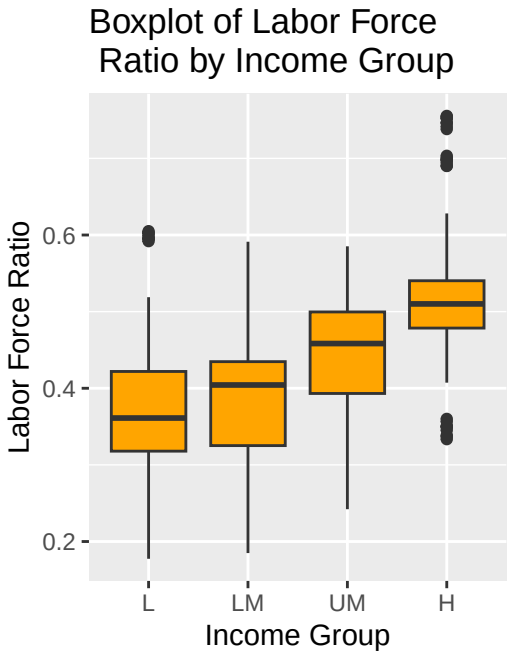
We started with some exploratory data analysis to understand the impact of life expectancy as a function of adjusted net national income and notice that across various income groups, there is a positive correlation between the net national income of a country and the life expectancy of its

people, illustrated in the figure below. This relationship highlights the significant role economic resources play in improving the quality of life and access to essential services, such as healthcare, nutrition, and education, which collectively contribute to increased life expectancy. We also evaluate the model by including an interaction term between labor force and life expectancy, considering that for certain (higher) values of life expectancy, we could expect to see an increase in labor force.



We further examine the relationship between the countries classified under various income groups and the corresponding life expectancy and labor force ratio for a given year. We again note that countries classified within higher income group brackets tend to have higher labor force ratios, underscoring the critical role of an active labor force in driving economic growth and maintaining a higher income classification.





We run the proportional odds model to try and model the income group classification of a country and to understand how accurately our model performs. The intercepts are detailed in the table below:

Term	Estimate	SE	t-value	p-value
Life Expectancy (Years)	-0.24	0.02	-16.04	<0.001
Labor Force Ratio	-88.98	0.33	-265.79	<0.001
Intersection Term	1.39	0.01	103.21	<0.001
L/LM	-16.26	0.86	-18.88	<0.001
LM/UM	-13.06	0.95	-13.66	<0.001
UM/H	-10.05	1.01	-9.87	<0.001

We notice the statistical significance of both the life expectancy and labor force ratio terms along with the intercepts for boundaries between the various levels of the income group are statistically significant with an overall accuracy of 66.4% (95% confidence interval of 63.64% - 69.1%) and a no information rate (NIR) of 33.9%. The confusion matrix from the testing data have been shown below. We do notice that there seems to be a slight difficulty in classification between Lower Middle (LM) and Upper Middle (UM) levels which can be improved with more extensive training data.

Prediction	L	LM	UM	H
L	80	53	13	0
LM	77	182	59	0
UM	0	71	184	67
H	0	7	44	328

We used the Brant test to evaluate the validity of the proportional odds assumption. With the result Brandt model, we notice that the proportional odds assumption does not hold hence we

can not use the ordinal linear regression model. We then tried to model the income group of the country based on a multinomial regression model adding an intersection term of Life Expectancy and the Labor Force ratio of the country, the results are detailed below (Low Income group as the reference level):

Coefficients	Intercept	Life Expectancy (Years)	Labor Force Ratio	Interaction Term
LM	2.59	2.70	-1.79	-3.23
UM	0.58	1.12	3.56	4.71
U	-6.63	6.03	13.31	3.71

We note the Multinomial Regression model gives us 62% accuracy (95% confidence interval between 59%-64%) when we include the interaction term between the life expectancy and labor force ratio, along with a no information rate (NIR) 31.2%. This indicates our model does a reasonable job of predicting the income group status of a country for a given year. Compared to the baseline level (Low Income Group), we note that from the result of the multinomial model, for the Low Middle Income group the average life expectancy of citizens of the country play a bigger role than the labor force ratio, which is the inverse for the Upper Middle Income group where labor force ratio plays the bigger role. Even for the High Income group countries, labor force ratio is the biggest determining factor. The confusion matrix from the multinomial model is detailed below:

Prediction	L	LM	UM	H
L	126	64	15	0
LM	103	205	103	1
UM	8	122	205	72
H	0	1	74	393

Conclusion

The project aimed to uncover the drivers of global CO_2 emissions and explore the relationship between socioeconomic indicators and a country's income classification. Using the World Development Indicators (WDI) dataset, we analyzed variables such as population size, income levels, education years, and urbanization rates to address two key questions. This study offers valuable insights into the global dynamics of CO_2 emissions and their link to socioeconomic indicators. By analyzing comprehensive data from the World Bank, we investigated the impact of variables like population size, income, and education on both emissions and income classification.

Through a systematic process of data sourcing, preprocessing, variable selection, and rigorous statistical analysis—including linear and non-linear regression models—we identified meaningful patterns. According to the statistical analysis , we found that income and population size are the most influential predictors of CO_2 emissions, with wealthier and larger countries contributing disproportionately. Education, while less directly linked to emissions, plays a crucial role in how National Income impacts Carbon Emission. These results underscore the interplay between economic activity and environmental impact, highlighting the need for tailored strategies to address emissions across different income groups. These patterns show that solving problems like climate change and

poverty needs different plans for different countries—rich, high-emission countries need to lead the way in cutting emissions and sharing resources to help others grow in a fair and sustainable way.

We also see that life expectancy and labor force ratio are critical indicators of a population's health and economic activity, both of which are strongly tied to a country's income group classification. Higher life expectancy and a larger labor force ratio are often associated with wealthier, more developed economies, while lower values indicate the challenges faced by poorer regions. The model highlights the importance of these factors in understanding socio-economic development, while also acknowledging that classifying income groups is complex and requires further refinement, especially when distinguishing between the middle income groups.

Future research should integrate industry-specific data, such as emissions from energy and transportation sectors, and leverage advanced modeling techniques to deepen understanding. This could guide policy interventions aimed at balancing economic development with environmental sustainability.

References

[1]World Development Indicators (2024). World Bank Group (Databank). Source Data. *(Link)*

[2]International Energy agency (February 2023). "The world's top 1% of emitters produce over 1000 times more CO_2 than the bottom 1%". *(Link)*

[3]John Creamer (September 2024). United States Census Bureau. "Health Inclusive Poverty Measure in the United States: 2023". *(Link)*

[4]Our World in data. "CO_2 emissions per capita vs GDP per capita, 2022". *(Link)*

[5]National Library of Medicine, 2016. "The Association Between Income and Life Expectancy in the United States, 2001–2014".*([Link]https://pmc.ncbi.nlm.nih.gov/articles/PMC4866586/)*