Environmental Research Report

Analysis of Worldwide Carbon Emission Contributions

Suzy Anil, Sukhpreet Sahota, Xianchi Zhang, Yuanjing Zhu

2022-12-02

To Michael S. Regan and global environmental agency administrators/directors,

Abstract:

During the duration of our study, our group focused on better understanding and modeling economic factors into a nation's carbon emissions. Our analysis below highlights this objective in a two-folded manner, centered on understanding how different economic status groups affect a nation's carbon footprint as well as using key economic and demographic characteristics to categorize nations into high, medium, or low carbon emitters. This categorization would enable countries to understand how they compare to other countries around the globe.

To conduct this study, we used the World Inequality Database as our primary source of data to obtain the income bracket distribution, the national wealth, the national income, population, and the nation's carbon footprint from ten countries geographically spread around the world. As a result, we were able to conclude the following:

1. The amount of carbon emissions was not the same across all defined income brackets as the bottom 50% accounted for the most carbon emission within a country.
2. By analyzing the logarithmic value of a nation's income, wealth, and population, we have derived a model that predicts and categorizes a country as a high, medium, or low emitter with 95% accuracy.

Introduction

When one turns on the news and sees natural disasters caused by weather, the next topic most meteorologists will expose viewers to is climate change. This global phenomenon is defined as the shifts within temperatures and weather patterns caused by human activities, which lead to significant climatic events. With more and more significant climatic events on the rise, more and more businesses and countries have changed their perspective and level of effort to combat this in the near future (outlining the year 2030 as the target for ensuring and implementing climate initiatives). For example, companies, such as Apple, have developed more sustainable and climate-conscious global supply chains and operations that aim to be carbon neutral. Similarly, many global countries and leaders have stepped forth to show the importance of this matter through initiatives/agreements such as the Paris Agreement/Paris Accords and establishing other global and national environmental targets.

In conjunction with this shift in climate, there is also a pronounced change within the distribution of wealth within countries across the globe. As the world has adopted technological advancements, the number of billionaires within the world have also increased dramatically. In 2021 alone, the world saw 153 new billionaires – an astounding 3 new billionaires per week (Block, 2022). The creation of this new wealth has only led to an increase in economic disparity within the world. The International Monetary Fund (IMF)

discovered that "the current disparities are extreme. The poorest half of the global population owns just €2,900 (in purchasing power parity) per adult, while the top 10 percent owns roughly 190 times as much. Income inequalities are not much better. The richest 10 percent today snap up 52 percent of all income. The poorest half get just 8.5 percent" (Staley, 2022). As seen in the graphic from The IMF in the Appendix (titled *A lopsided world*), the Fund also concluded from their analysis that 48% of global carbon emissions are caused by the top 10%. Using this as a baseline, we wanted to understand if there was truly a difference on the level of carbon emission between income levels.

To conduct our study, as aforementioned, our dataset was selected from the World Inequality Database (WID), which is one of the most extensive databases on the evolution of world distribution of income and wealth within and between countries. The database is open-access and has compiled valid data from national databases, surveys, fiscal data, and wealth rankings. With its vast array of features, there are many key economic and social inequality questions that could be answered with access to this data. Our group has decided to focus our statistical analysis on the impact of certain economical features on a nation's carbon footprint (Total National CO_2 Footprint). For clarification, a nation's total carbon footprint is equal to the combination of CO_2 footprint and footprint of other greenhouse gases.

While the dataset/database is vast, we narrowed down our analysis to the following key variables that will help us effectively analyze and assess the impact of economic and demographic statistics on carbon emissions for a subset of ten select countries (The United States, China, India, Germany, the United Kingdom, Canada, Australia, Brazil, Nigeria, and South Africa): national income, GDP, income inequality, population, market-value national wealth, years (from 2000 – 2020). *It is important to note: to help standardize the findings for all countries, the US dollar was the currency selected for the appropriate variables*

This leads to two distinct research questions:

1. How do income brackets (top 10%, middle 40%, bottom 50%) affect a nation's carbon footprint? Based on the research mentioned above as well as additional publications/evidence highlighting that the wealthiest bracket emits notably more tons of carbon compared to the bottom bracket (Ritchie, 2018), we hope to use our data and our model to either validate this evidence or understand the true relationship between the different economic status groups and their respective effects on a nation's carbon footprint. Our initial assumption refutes the evidence highlighted and is that carbon emissions are comparable amongst the three income brackets.

a. H_0: Carbon emissions are the same across all income brackets.
b. H_A: Carbon emissions vary across income brackets.

If we see a disparity between the three income brackets, we also want to understand the income bracket that emits the most carbon. Is it truly the wealthiest top 10%?

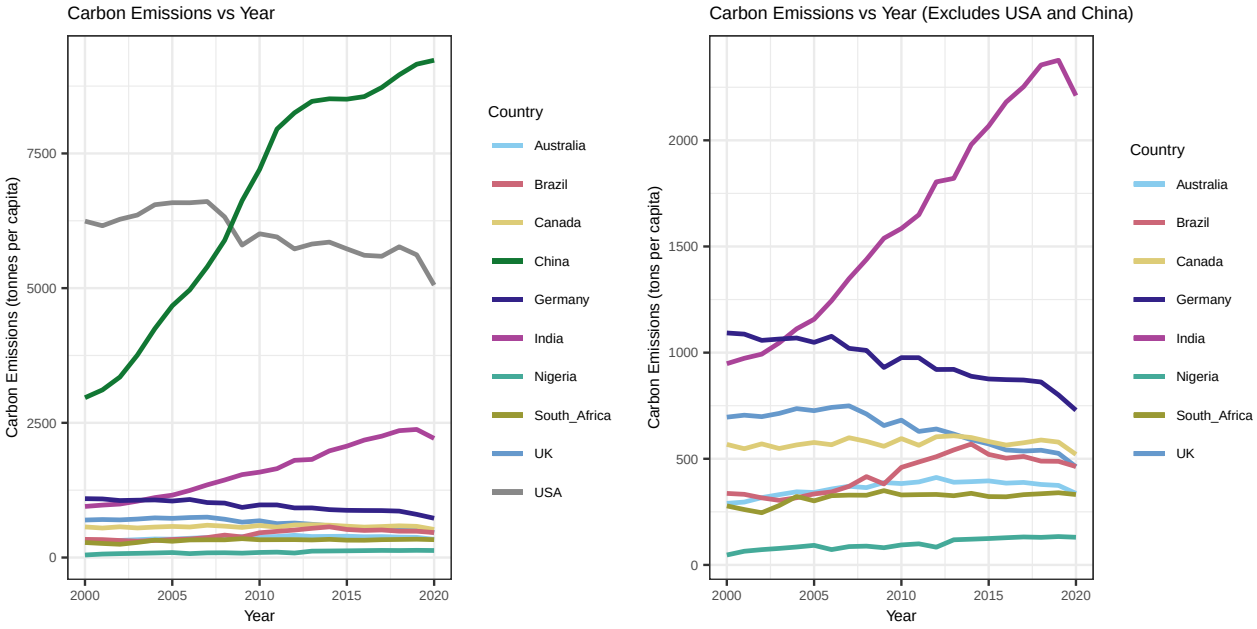
2. Given a nation's economic and demographic statistics, how accurately is an environmentalist able to classify/categorize countries into their appropriate level of carbon emissions? Our hopeful end goal is to create and use a model to predict the appropriate carbon emission category for other countries (unseen data). One of our key assumptions in creating this model is that there are no CO_2 changes between any of the income brackets and changes within a country's average wealth over time have not had an impact on the country's respective carbon footprint, and thus, the difference in carbon emission class is proportional.

Methods

Data Exploration

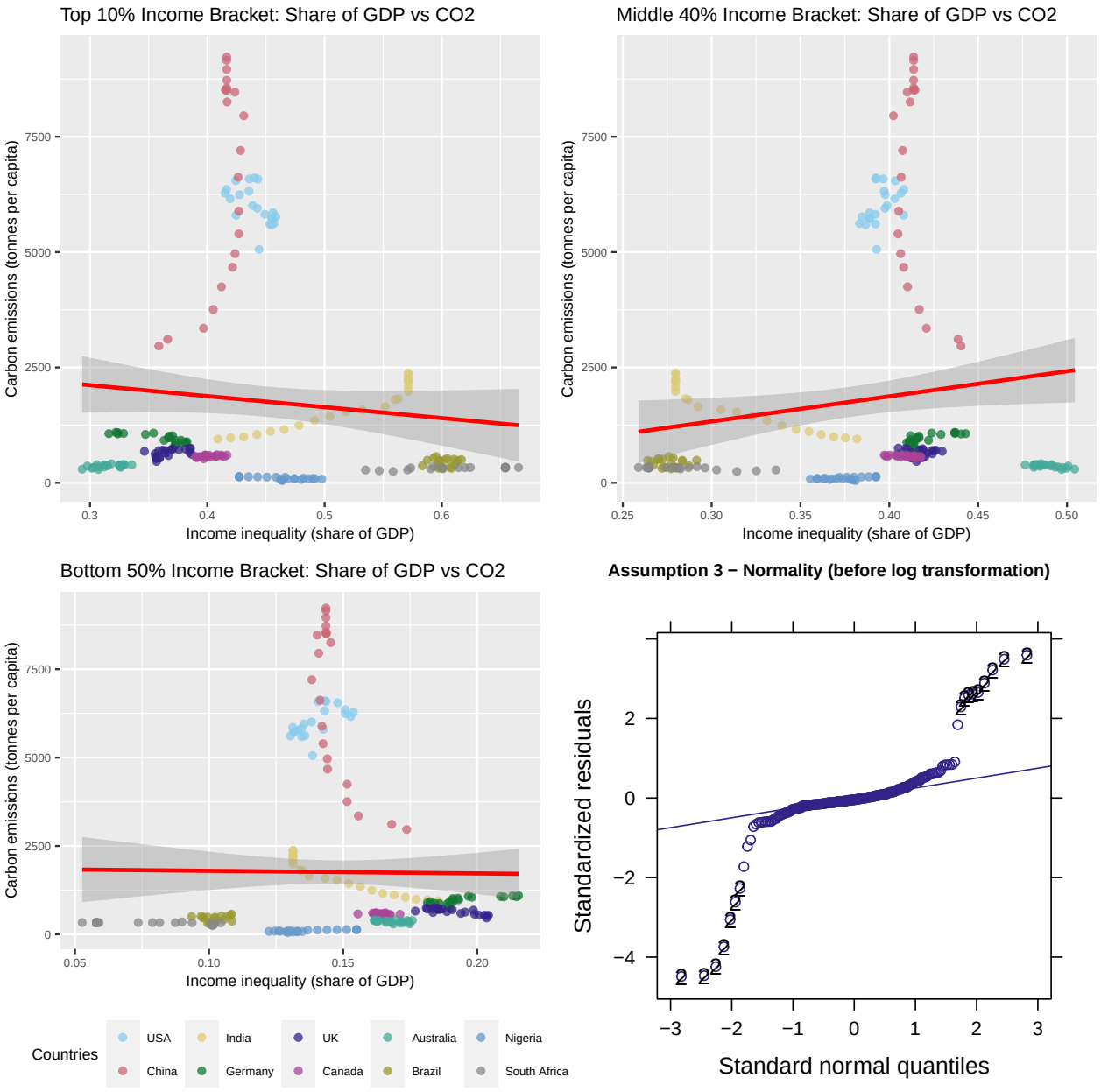
To begin our analysis, we explored each country's carbon emission as well as the relationship between that response and the other key economic indicators mentioned above. From the following graphs, we plotted CO_2

over time for every country. In the graph to the left, we included all ten countries and realized a disparity in trends between China and USA versus the other eight. Thus, we classify USA and China as high emitters. The graph on the right zooms in on the remaining eight countries, which we classify as medium/low emitters. There is a linear trend for all countries, which allows us to assume linearity for the models we will fit.

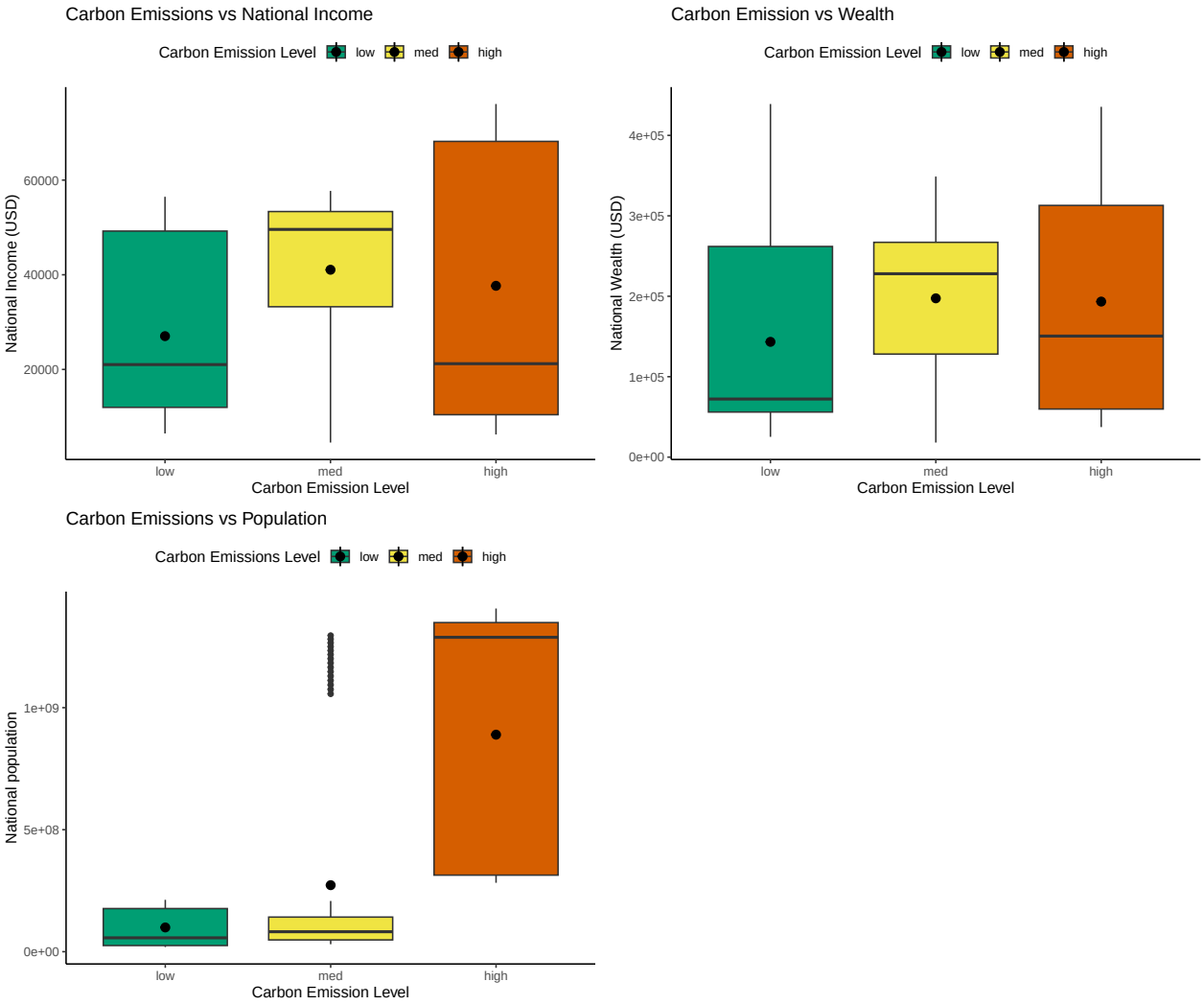


Upon initially recognizing this linear relationship, we took a deeper dive and analyzed each country separately. From the series of graphs found within the Appendix (titled *Relationship between GDP vs CO_2*) to see if there was a general trend between GDP and CO_2 shared between all the countries, we found that there is no common direction. Since the graphs show the differing relationships between GDP and CO_2, we would not be able to build a generalizable model that could apply to countries unseen by the model and thus opted not to use GDP as our indicator for national income/wealth. During our initial exploratory analysis, we discovered most countries show positive correlation between wealth and CO_2 (except for USA, Germany, and UK). This finding was crucial for model development as the model needs to account for the differences between countries.

To analyze the distribution and impact of income brackets, we fit a line to each of the above graphs and saw a clear distinction between the high and low emitters. The high emitters, China and USA, rise high above the fit line and the rest either sit on the line or below. This is consistent across the three graphs. To correct for the disparity between the high and low emitters, we need to consider a linear transformation on carbon emissions to create a best fit model. Additionally, our model did not pass the normality assumption, further validating the impact and variance between the high emitting countries vs medium/low emitting countries and further validating that a transformation is required.



Citing the IMF, the predictors that are important in properly categorizing the level of carbon emissions for a respective country to include within our best fit model are national income, market-value national wealth, and population size. Per the WID and our trend analysis of GDP, national income is the only income concept that has an internationally agreed definition (established by the United Nations System of National Accounts, see SNA 2008). Similarly, the national economy/wealth - in the national accounts sense – encompasses all domestic sectors, i.e. all entities that are resident of a given country (in the sense of their economic activity), whether they belong to the private sector, the corporate sector, the government sector. Lastly, to account for the disparities between our select countries, it is best to include population. While the WID has additional variables, the variation amongst countries for those respective variables is too country-specific that it would inhibit our ability to create a generalized model. To ensure that our model did not have any missing data, was complete, and representative, we subsetted and selected countries from WID with data available across all aforementioned predictors across all 20 years (2000 – 2020).



Data Modeling and Assessment

To account for the differences between countries, the best fit models for both of our research questions are mixed models. Similar to linear regression models, mixed models will enable us to better analyze the carbon emission while estimating the random effect/impact of each country. In addition to taking the impact of country into account, a mixed model will also take into account any correlation that exists within our data. The generalized linear mixed model can be represented through the following framework/model formula:

$$Y = X_i\beta + Z_j u + \epsilon$$

Question 1: With this initial exploratory data analysis serving as a backdrop, we decided to log transform our outcome variable, carbon emissions, to account for the high variance between high and low emitters (which can be seen through the aforementioned Income Bracket: Share of GDP vs Carbon Emissions graphs above). We opted to use a linear mixed model as compared to a linear model and included country as a random effect to make our model more generalizable since we found each country followed slightly different trends.

The i and j within the above generalized linear mixed model equal 3 (for the number of income brackets) and 10 (for the number of countries) respectively to represent this research question. This equates to the

following model:

$$Log(CO_2) = \beta_1 Top10 + \beta_2 Middle40 + \beta_3 Bottom50 + (1|Country)$$

For this generalized model, we quickly assumed that the similarity amongst the predictors (all three income brackets) would lead to multicollinearity. We were able to assume this since the shares for all three income brackets total 1. However, because the predictors are interdependent but independent from the response variable (carbon emissions), we can continue with our analysis under this assumption. We are able to resolve this multicollinearity issue within our model by including a negative intercept term, which will offset the intercept brought forth by these income predictors totaling 1 and by the random countries maintaining an intercept. We are able to confirm this resolution with VIFs being lower than 5.

Table 1: VIF Multicollinearity Check for Q1: CO2 and Income Bracket Analysis

top10	middle40	bottom50
1.68	3.76	4.88

Question 2: In the CO_2 vs Income and CO_2 vs Wealth graph, we do see major differences in range and mean among all the three carbon emitter levels. However, in the CO_2 vs population graph there is a major difference between low/medium emitters compared to high emitters. Logically, this makes sense, since with more people a country requires more resources and energy to accommodate them. Based on this finding, we applied a log transformation to all predictor variables (population, national income, and wealth) to normalize the differences amongst the countries.

Our goal is to predict each country's level of carbon emissions for each year. Since the outcome is ranked low, medium, and high, an ordinal model would be the best fit. An ordinal model operates on the proportional odds assumption which we checked with the Brant test (a test/function that calculates parallel regression and assesses proportionality amongst predictors) and found that the model does not violate this assumption (with the omnibus and the associated predicted having probabilities equal to or close to 1). We also checked for multicollinearity and found that all VIF values fell under the threshold.

Table 2: Proportional Odds Assumption Check for Question 2: Carbon Emissions and Economic/Demographic Indicators Analysis

	X2	df	probability
Omnibus	0.0002	3	1.00
income_log	0.0001	1	0.99
wealth_log	0.0001	1	0.99
pop_log	0.0002	1	0.99

Table 3: VIF Multicollinearity Check for Q2: Carbon Emissions and Economic/Demographic Indicators Analysis

income_log	wealth_log	pop_log
1.00	1.34	1.40

Since all assumptions have been held, we fit a cumulative link mixed model to predict the ordinal outcome variable. Similar to the generalized linear mixed model used to represent Question 1, a cumulative link

mixed model enables one to analyze ordinal response variables while still maintaining random effects. As we aim to categorize and rank countries based on their respective carbon emission levels, it is more powerful to maintain order as compared to a multinomial model. To maintain and account for this order, the generalized linear mixed model is tweaked to the following model:

$$Y = \alpha_j - X_i\beta + Z_{t\ [\ i\ ]}u_t + \epsilon$$

where α is the intercept/threshold coefficient between the different level comparison combination (i.e. Low | Medium and Medium | High), the i, j and t within the above cumulative link mixed model equal 3 (for the number of predictors (income, wealth, population)), 3 (for the number of emission levels (low, medium, high), and 10 (for the number of countries) respectively to represent this research question. This equates to the following model:

$$P(Y \leq j) = \alpha_j + \beta_1(LogIncome) + \beta_2(LogWealth) + \beta_3(LogPopulation) + (1|Country)$$

Results

Model 1: Linear Mixed Model As shown through the model output within the Appendix titled *Summary Results for Model 1*, our resulting model to test the equivalence amongst the income bracket is the following:

$$Log(CO_2) = 8.92Top10 + 2.89Middle40 + 10.74Bottom50 + Z_j u$$

With country as our random effect, we tested the fixed effects with a Wald test. The Wald test compares the coefficient's estimated value with the estimated standard error for the coefficient. Our null hypothesis states that the variance between coefficients for the random effect is zero. With a 5% significance level, we found all three predictors to be statistically significant. We can also confirm there is no variance between the coefficients of the three income brackets when inferring their relationship with carbon emissions.

Table 4: Model 1: Marginal and Conditional R-squared

R2m	R2c
0.06	0.98

When analyzing the model for how representative it is, the R_2 outlines this but in two distinct measures: marginal R-squared and Conditional R-squared. Marginal R_2 is concerned with the variance explained by fixed effects. Our model returned low marginal R_2 (6.3%) which indicates the fixed effects do not explain the variance in the outcome.

On the other hand, conditional R-squared is concerned with the variance explained by both fixed and random factors of the entire model. We have a very high conditional R_2 that accounts for 98% of the variance in the outcome variable which indicates country determines a lot of the outcome variance. Random effects explain additional variance compared to the fixed effects. With just a linear regression model, we still see a very large overall R_2 (0.97).

The last verification to understand how representative this model truly is and whether it can be considered a valid model for answering our research question are the three assumption graphs. After including country, the model passes the normality assumption as well as other model assumption of homoscedasticity and linearity, unlike previous attempts without country as a random effect (see graphs within the Appendix titled *Model 1 Assumption Check*).

Model 2: Cumulative Link Mixed Model Our resulting model (derived from the model output within the Appendix titled *Summary Results for Model 2: Cumulative Link Model*) to classify a country's carbon emission is the following:

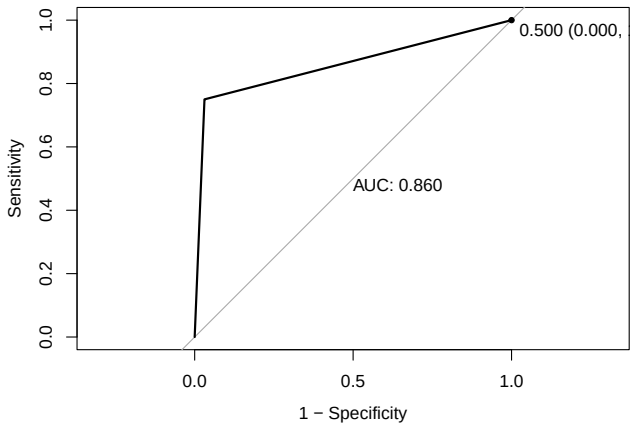
$$P(Low \leq Medium/Low > Medium) = 601.38 + 6.82Log(Income) + 12.15Log(Wealth) +$$
$$21.25Log(Population) + Z_{t\ [\ i\ ]}u_t$$

With our specified standards for high (> 2000 tons), medium (500 – 2000 tons), and low (< 500 tons) emitters, there are only two countries in the world that fit the definition of a high emitter: USA and China. Since we trained our model with both countries, we do not have other countries to test our model's ability to predict high emitters. The disparity between high and low/medium emitters is so large, we are more concerned about the precision of our model distinguishing between low and medium emitters; therefore, for prediction, we decided to limit our test set. For our test set, we have selected five countries (Argentina, Ghana, Mexico, Norway, and Qatar) which are similar to our training set in that they are geographically spread out and fall between these two categories. As a result, as can be seen within our prediction, our model predicted their emission levels with 95% accuracy which is greater than the no information rate. As seen through our confusion matrix, there is high sensitivity with the low class, meaning our model correctly classified most low emitters (approximately 96%). In comparison, our model correctly classified 75% of countries as medium class carbon emitters; however, as is evident through the confusion matrix, the number of years countries distinguished as low compared to medium is more than 12 fold (97 to 8). Additionally, The ROC curve has an AUC of 0.86 (which is 0.14 units away from the ideal value of 1), further validating our model does very well in predicting low and medium carbon emitters over the years.

Running predictions on a smaller test set:

Table 5: Confusion Matrix - Predicting CO2 Level Classification

	Reference low	Reference med
Pred low	94	6
Pred med	3	0



Conclusion:

Through this limited study, we were able to ultimately refute the conclusion presented by the IMF as we did not find the top 10% income bracket contributing the most to the level of carbon emissions within a country, but rather the bottom 50%. While concluding this, we do note that our select and subsetted data analysis may lead to differing results and should be expanded to include more countries, more factors, and

more models to see if our conclusion stands or gets closer to the one presented by IMF. As we assessed our prediction results, we were able to conclude that the high accuracy provides evidence that national income, national wealth, and population are good economic and demographic indicators for determining a nation's carbon emissions.

While our personal ambitions were to compose a comprehensive study, we know that limitations within our dataset, our model, and current available R packages lend to us falling short on our thoroughness. As detailed within the Methods section, it would be ideal to incorporate more descriptive country feature/statistics/information, such as factors other than the economy: energy consumption, or foreign investment. These additional variables would provide more context to our model and allow for a better comparison to the work studied by IMF. In regard to Question 2, the variation amongst the low, medium, and high levels is not apparent. It would be best to understand if there is a better measure to define these levels to create distinct class, and thus variation amongst the levels. We also found this difference in carbon emission levels by testing our model on the same number of countries as the number of countries within our training data (i.e. initially selected 10 countries). However, these attempts showed accuracy results that were poor due to our sample size being small.

To address these limitations, our future work would consist of the following:

1. Adding supplementary descriptive statistics/feature on a country. We would like to incorporate data on the government sector vs private sector as well as more information on personal wealth to better define the characteristics of each income bracket
2. Adding more countries with carbon emissions that were not represented within our model (i.e. extremely low carbon emission rates). These additional training countries would enable us to test our model on a larger test size.
3. Rather than just using mean, we would want to define variance based on a weighted average and use this information to define unique levels of carbon emissions. We would also supplement our test data with synthetic data to replicate/incorporate countries with high carbon emissions as we noticed that there were none outside the US, China, and India. Currently, we have the 3 levels of carbon emissions, but this measure would possibly create more levels (i.e. extremely low, low, medium, high, extremely high).
4. Lastly, it would be best to understand how this research can play a role in the big picture by combining our interpretations different governmental policies and national context.

References

Block, F. (2022, March 17). The world's billionaire population tops 3,381, adding 3 each week last year. Barron's. Retrieved November 26, 2022, from https://www.barrons.com/articles/the-worlds-billionaire-population-tops-3-381-adding-3-each-week-last-year-01647548271

Ritchie, H. (2018, October 16). Global inequalities in CO2 emissions. Our World in Data. Retrieved November 4, 2022, from https://ourworldindata.org/co2-by-income-region

Staley, A. (2022, March). Global inequalities. IMF. Retrieved November 4, 2022, from https://www.imf.org/en/Publications/fandd/issues/2022/03/Global-inequalities-Stanley

Appendix

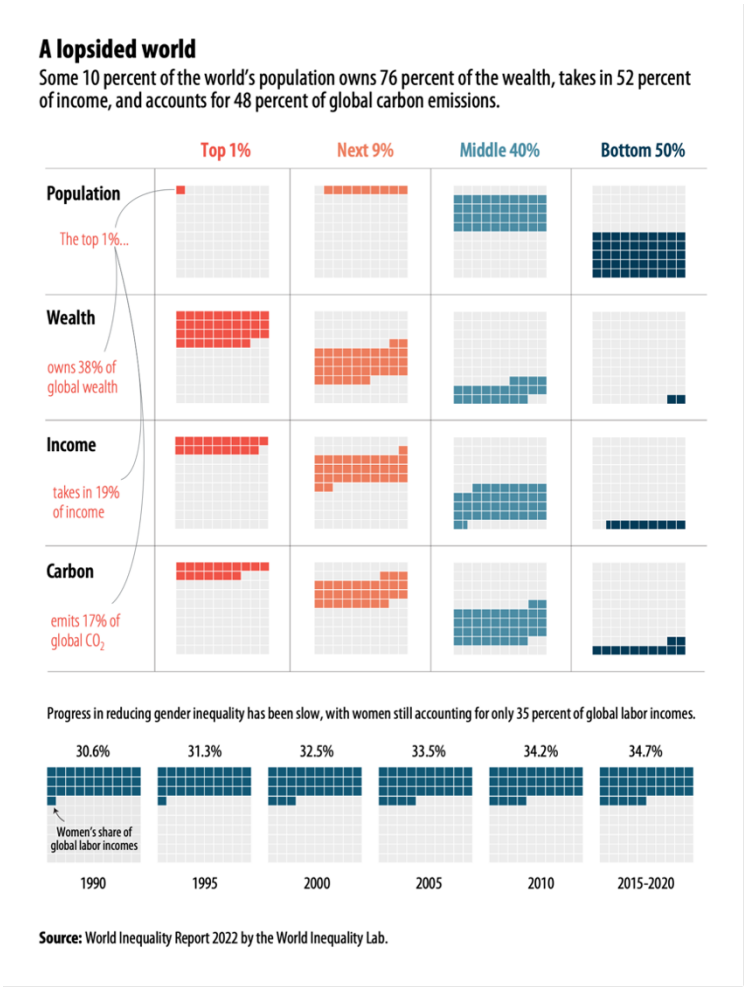


Figure 1: an image caption Source: A lopsided world

Figure 2: Relationship between GDP vs CO_2

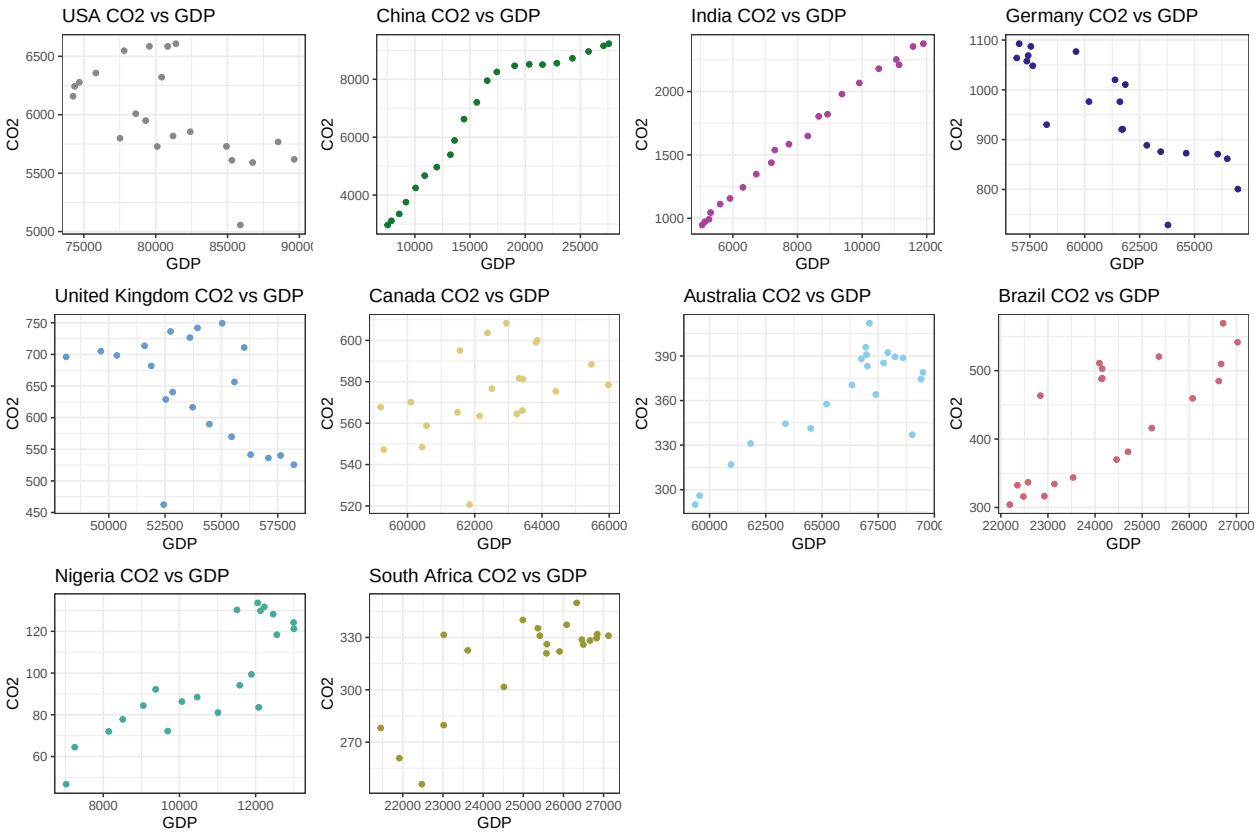
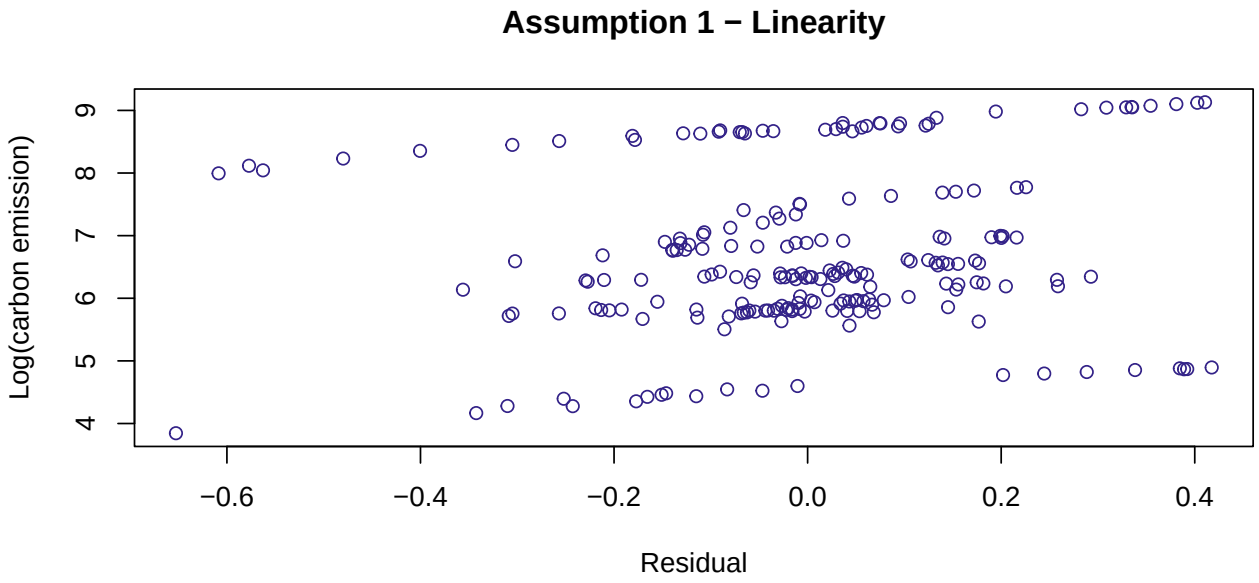


Table 6: Summary Results for Model 1

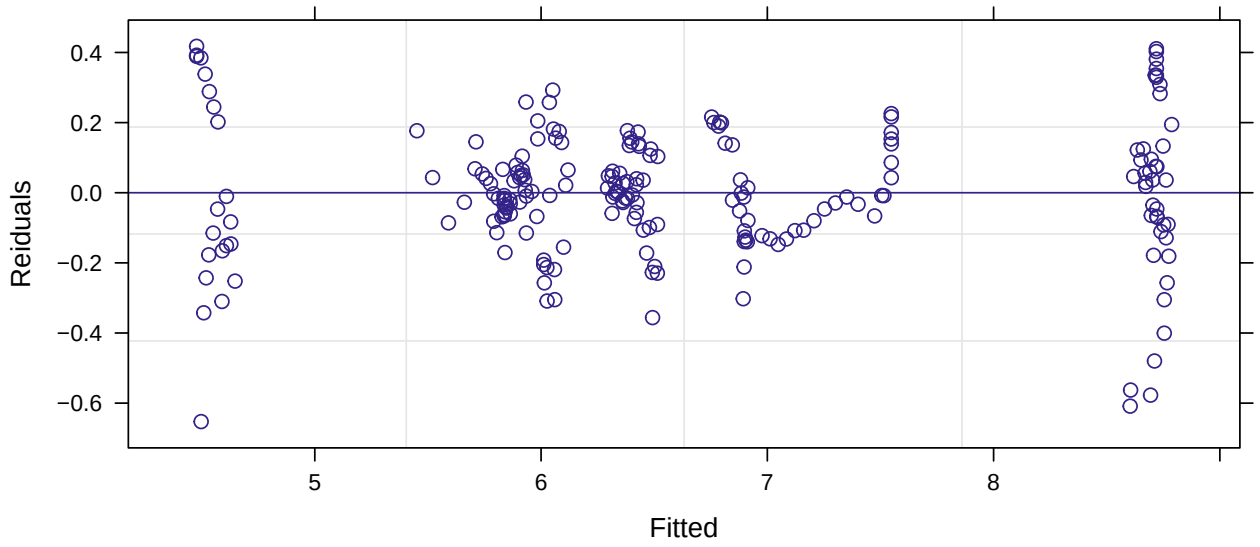
	log(stacking_co2)
top10	8.92***
	p = 0.00
middle40	2.89**
	p = 0.04
bottom50	10.74***
	p = 0.0001
N	210
Log Likelihood	19.88
AIC	−29.75
BIC	−13.02

***p < .01; **p < .05; *p < .1

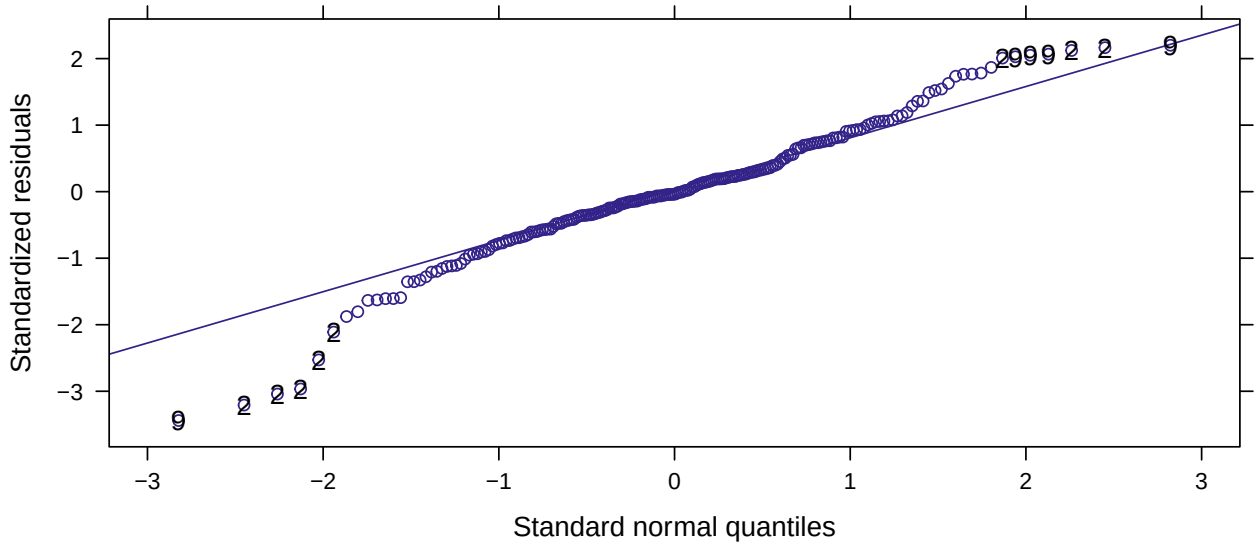
Figure 3: Model 1 Assumption Check



Assumption 2 − Homogeneity of Variance/Homoscedasticity



Assumption 3 − Normality (after log transformation)



	Summary Results for Model 2: Cumulative Link Model
low|med	601.379
	(0.001)
med|high	627.388
	(0.001)
income_log	6.817
	(0.002)
wealth_log	12.149
	(0.002)
pop_log	21.252
	(0.003)
SD (Intercept Country)	11.782
Num.Obs.	210
R2 Marg.	0.790
R2 Cond.	0.995
AIC	62.4
BIC	82.5
RMSE	1.17