Relationship between carbon emissions and development

Anjali Agrawal and Eileen Zhang

10 May 2022

1 Topic and Explanation

In recent years we have seen an exponential increase in the concern for our climate. The consequences of environmental decay are affecting mundane lives in an affronting way that has made climate action imperative. To explore the nature of this change, we decided to investigate if there was a relationship between CO2 emissions per capita growth and GDP per capita growth. Our theory is that increasing levels of growth in CO2 emissions per capita signals a higher level of development for a country. We used data from the World Bank spanning 185 countries from 2000 to 2019 and ran machine learning algorithms to present our analysis. We first began by running an Ordinary Least Squares regression on the data. From there we ran an endogeneity falsification test on our data by swapping the dependency of variables. We decided to test the Environmental Kuznets theory while we had our variables swapped. Additionally, we created dummy variables for Low, Middle, and High income levels (as assigned by the World Bank) and ran a dummy regression. We found that adding interaction terms to this regression improved the fit of our model. Finally, we ran a panel regression on the main greenhouse gases: CO2, Methane, and NOx, with time fixed effects and country fixed effects.

2 Files we included

In the file **Programmatic data download.ipynb**, we used the *wbgapi* API that accesses the World Bank database. With this file, users can programmatically download data frame that we created. Using this exported data, users can then run the file **Data analysis.ipynb**. With this file, users can run the analysis that we did i.e. all our regressions and also programmatically download our visualisations to reproduce our results.

3 Analysis and Methods

3.1 Impact of carbon emission on GDP

For our initial regression we ran an OLS linear regression on log GDP and log CO2. We did this by extracting the CO2 emissions per capita and GDP per capita variables and logged the data. We then dropped the NaN values from the data and filtered the time period to be between 2000 to 2019. With the log GDP per capita as our dependent variable, our model achieved an R^2 value of 0.743 (as seen in Figure 1.). The following four scatter plots give us a visual relationship between our variables. In the first image we can see that the majority of our data points lie within the black segment which represents the standard deviation from the slope. In this case, our R^2 value indicates that 74.3% of our data should lie within this region and any remaining values are not adequately explained by this model. The residual versus log GDP scatter shows the residuals on the vertical axis and GDP per capita growth on the y-axis. Residuals are a measure of how well the data fits the curve; the closer the data point is from the horizontal line at 0 the better our model does.

$$log\_GDP_{it} = \alpha + \beta log\_CO2_{it} + \epsilon_{it} \tag{1}$$

Figure 1: Fitted explanations of Regression 1

	log_gdp
const	7.76***
	(0.01)
log_co2	0.86***
	(0.01)
R-squared	0.72
R-squared Adj.	0.72

Table 1

3.2 Panel Regression

$$log\_GDP_{it} = \alpha + \beta_1 log\_CO2_{it} + \beta_2 log\_Methane_{it} + log\_NOX_{it} + \beta_i + \beta_t + \epsilon_{it} \tag{2}$$

We created a panel dataset that introduced methane emission growth and nitrous oxide emission growth. We controlled for time fixed effects and also country fixed effects (in the above equation β_i and β_t represent the vector of coefficients for fixed and time effects) which add additional year and country dummy variables to our regression. This should help us remove any omitted variable bias from our data set resulting from differences between years or countries. This resulted in a significant decrease in the fit of our data compared to the OLS linear regression with an overall R^2 of 0.2536 as can be seen in Table 2. This also confirms when we see Figure 2, where we can see that by adding in multiple controls, our data faces multicollinearity and we had initially overestimated the effect of carbon di oxide emission growth on GDP growth.

4 Extension

4.1 Dummy regression and Environmental Kuznets curve

As a falsification test, to check the validity of our results, we flipped the dependency of our variables to check for endogeneity. We found that this OLS linear regression was also significant and therefore confirmed our suspicions of simultaneous causality. Given that we were now analysing the relationship between growth of GDP on growth of CO2 we decided to analyse the environmental Kuznets curve with our data. The environmental Kuznets curve (EKC) is a hypothesized relationship between various indicators of environmental degradation and per capita income. In the early stages of economic growth, pollution emissions increase and environmental quality declines, but beyond some level of per capita income (which will vary for different indicators) the trend reverses, so that at high income levels, economic growth leads to environmental improvement. This implies that environmental impacts or emissions per capita are an inverted U-shaped function of per capita income In order to investigate

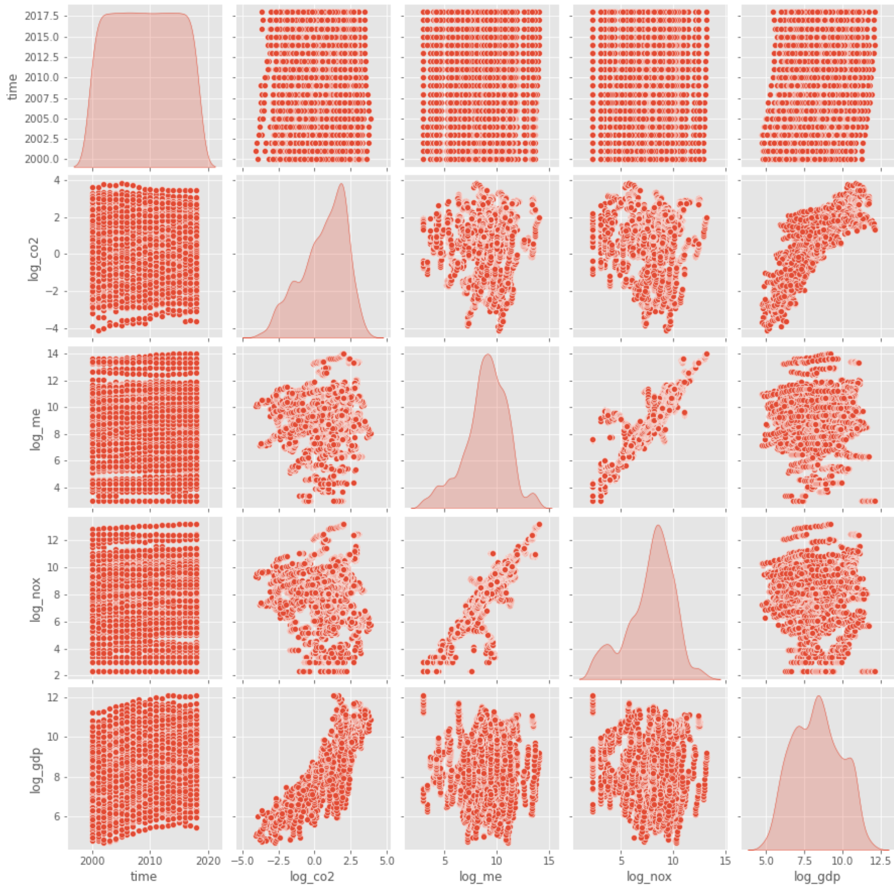

Figure 2: The graph shows pair plots of the variables we used in our panel regression. The plots here show density functions and scatter plots to represent multiple correlations within our explanatory variables.

Dep. Variable:	log_gdp	**R-squared:**	0.2108
Estimator:	PanelOLS	**R-squared (Between):**	0.2334
No. Observations:	5321	**R-squared (Within):**	0.2198
Date:	Wed, May 04 2022	**R-squared (Overall):**	0.2536
Time:	20:37:41	**Log-likelihood**	629.25
Cov. Estimator:	Unadjusted		
		F-statistic:	455.41
Entities:	185	**P-value**	0.0000
Avg Obs:	28.762	**Distribution:**	F(3,5115)
Min Obs:	3.0000		
Max Obs:	38.000	**F-statistic (robust):**	455.41
		P-value	0.0000
Time periods:	19	**Distribution:**	F(3,5115)
Avg Obs:	280.05		
Min Obs:	269.00		
Max Obs:	289.00		

	Parameter	**Std. Err.**	**T-stat**	**P-value**	**Lower CI**	**Upper CI**
const	4.7730	0.2082	22.921	0.0000	4.3648	5.1813
log_co2	0.4261	0.0165	25.863	0.0000	0.3938	0.4584
log_me	0.2725	0.0250	10.890	0.0000	0.2235	0.3216
log_nox	0.0998	0.0264	3.7748	0.0002	0.0480	0.1517

Table 2: PanelOLS Estimation Summary

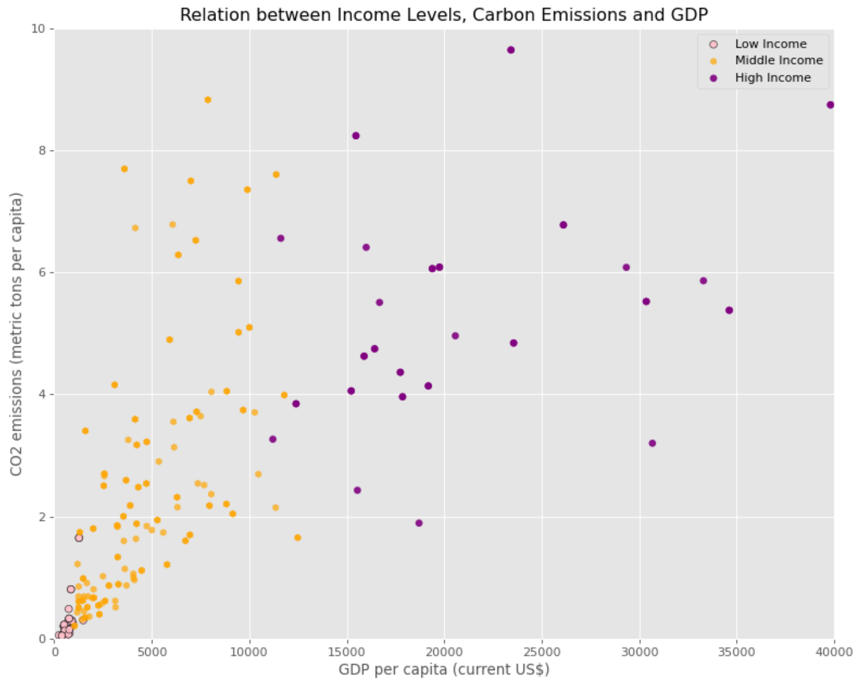

Figure 3: The colored dots represent each country in the data set. The pink dots represent low income countries, orange dots represent middle income countries and purple dots represent high income countries as classified by World Bank. To avoid clustering, we display only the observations of year 2008. We can see a vague resemblance to the Kuznet Curve as some high income countries choose to have low emissions

whether we could find a hint of a similar relationship we created income level dummy variables. According to the World Bank, we assigned countries into low income, middle income, and high income categories. We did this by first adding countries into their respective lists. Then using helper functions we assigned the country with a 1 if they were in the income level list and 0 if not. We modelled the following model:

$$log\_C02_{it} = \alpha + \beta_1 log\_GDP_{it} + \beta_2 Low\_Income_{it} + \beta_3 High\_Income_{it} + \epsilon_{it} \tag{3}$$

This improved the R^2 of our model up to 0.775. Looking at our results in Table 3, we can see that for high income countries the CO2 emissions per capita growth increases compared to other income level countries.

	log_co2
const	-6.23***
	(0.08)
log_gdp	0.70***
	(0.01)
Middle_dummy	1.24***
	(0.04)
High_dummy	1.28***
	(0.06)
R-squared	0.77
R-squared Adj.	0.77

Table 3

5 Interaction Terms

We hoped to improve the fit of our data even more by adding interaction terms to the regression. We saw in the previous results that each income level group had a different relationship (coefficient) with the CO2 per capita emission growth. An interaction term emphasises that the effect of variable that forms the interaction depends on the level of the other variable in the interaction. This means that by adding interaction terms we are allowing for each income level group its own independent relationship to the outcome variable thereby producing a more cohesive model. We created these terms by multiplying the log GDP with each dummy variable. Then we plugged these values into a new regression as follows:

$$log\_C02_{it} = \alpha + \beta_1 log\_GDP_{it} + \beta_2 Low\_Income_{it} + \beta_3 High\_Income_{it} + \beta_4 Low\_Income_{it} * log\_GDP + \beta_5 Low\_Income_{it} * log\_GDP_{it} + \epsilon_{it} \tag{4}$$

Adding these terms improved the fit of our model to an R^2 of 0.792. With each increase in 1% of the growth rate we see a 0.46% decrease in CO2 emissions per capita compared to middle income countries. This means that the CO2 per capita emissions are growing at a decreasing rate which suggests that there is some weight to the Kuznets argument. This strengthened our conclusion that higher income increases the level of pollution at a decreasing rate.

	log_co2
const	-5.74***
	(0.11)
log_gdp	0.80***
	(0.01)
Low_dummy	-2.99***
	(0.26)
High_dummy	4.53***
	(0.25)
Interaction_low	0.30***
	(0.04)
Interaction_high	**-0.47*****
	(0.03)
R-squared	0.79
R-squared Adj.	0.79

Table 4

6 Conclusion

We found that overall as we added more terms to our regression we were able to incrementally improve the fit of our data. Our variables of interest do suffer from endogeneity bias which weakens the results of our investigation but allowed for us to offer some insight into the validity of the EKC. We were able to offer some evidence that the EKC could offer valuable insight into the relationship between environmental degradation and GDP per capita growth. But with our panel dataset we saw that the R^2 value was significantly reduced which is a reflection of the biases that we have in our dataset that resulted in the over-estimation of causal effect.