Data Analysis 4 (advanced) - Assignment 2: To what extent does economic activity cause CO_2 emission?

Mihaly Hanics
Central European University
Quellenstrasse 51, 1100 Vienna, Austria
e-mail: hanics_mihaly@student.ceu.edu

ABSTRACT

The increase in gross domestic product (GDP) allows for increased industrial investments, that may result in emitting more fossil fuels, in turn, releasing more CO_2 into the atmosphere. With strong evidence that CO_2 emissions lead to destabilizing climate, it is important to find the mechanisms that cause increases in CO_2 emissions, and make international interventions to reduce these effects.

There is a strong positive correlation between high economic activity and CO_2 emissions. This work is a small example that showcases how one can investigate the extent of impact of a common economic indicator, GDP per capita, on CO_2 emissions (per capita) by looking at the yearly emissions and GDP data of 200 countries. The work investigates the causal relationship using multiple methods, and with the inclusion of controlling for country oil rents (an assumed confounder), showcasing how these methods can be used to attempt to infer the true impact of a variable on another. The analysis suggests that for a unit of increase in log GDP per capita (PPP, international $), there is an increase in more than half a unit of log total CO_2 per capita emissions (excluding LULUCF, Mt CO_2).

1 INTRODUCTION

Economic activity is strongly correlated with CO_2 emissions, a key variable of environmental impact (due to being the principal anthropogenic greenhouse gas affecting the Earth's radiative balance, causing anthropogenic climate change). It is therefore crucial for sustainable policymaking to find out the extent of this relationship, and whether it is causal.

In this work, I test 3 different models in overall 6 different settings to measure the impact of GDP per capita on total CO_2 per capita emissions, for 200 countries between the years 1992 and 2023. As a potential confounder, a measure for fossil fuel mining, particularly oil rent percentages (% of country GDP) are used as variable to control on in an extra setting for each model.

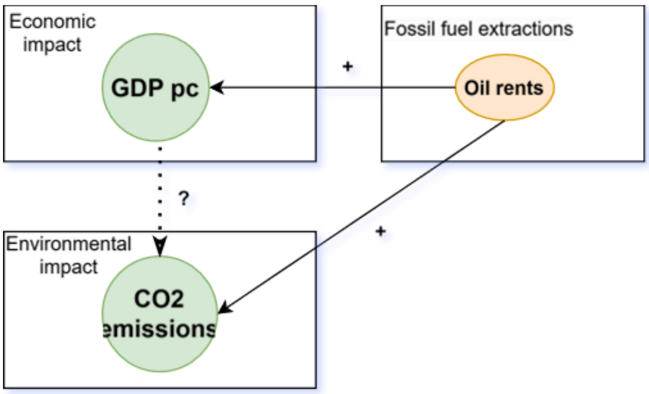
Code is available on GitHub

2 DATA

2.1 Data description

Data is collected from the World Development Indicators, from the World Bank Group [1]. The dataset consists of annual panel data covering multiple countries, with yearly assessment from 1992 to 2023. The key variables include GDP per capita (PPP, constant prices), CO_2 emissions per capita, and oil rents as a percentage of GDP. The inclusion of oil rents is motivated by its potential confounding role, given that resource-rich economies typically capture higher GDP per capita and CO_2 emissions per capita – countries such as Saudi Arabia may mislead calculations if not controlled for a variable that represents fossil fuel extraction. This could act as a “common cause” confounder (extraction of fossil fuels leads to higher economic activity, and also an increase in emissions), and therefore is a possible endogenous variable to control on. (Typically, oil extraction is a significant part of the GDP of countries that have oil reserves, making oil rents a relevant measure of fossil fuel induced economic development and environmental impact.)

The assumed causal relation is described by the following diagram:



In our models, the variable representing CO_2 emissions per capita in a country over a year is regressed on the variable representing GPD per capita in a country in a year.

In some settings, the variable representing oil rents (% of GDP) in a country in a year is controlled on. As this variable is not available for years 2022 and 2023 (for any country), such models use a less extensive dataset, which has impact on results, and have to be handled carefully.

2.2 Data preprocessing

There were multiple steps of preprocessing needed in order to ensure unbiased results, and to measure what we would like to measure.

Firstly, the variable for CO_2 emissions per capita needed to be constructed from country CO_2 emissions and population. Separate tables were created each measure, containing time-series data. Combined tables for running the models were also created.

The collected data included measures from the year 1960 to 2023, however we excluded any year's data that is before 1992. In the case of oil rents, oil rent data is not available in 2022 and 2023, therefore I excluded these two years' data and constructed a "restricted" table that includes measurements of all variables till 2021.

Since data is widely available, most countries hold data of all the variables in all included years. Therefore, I excluded the countries that have missing values of CO_2 emissions, population or GDP per capita from 1992 to 2023, and oil rents from 1992 to 2021. This excluded 66 countries (and economies) from the original listed 266, most of which are small nations, or nations not existent (anymore) as a separate country. Few more notable countries have been excluded as a result of missing data, including Sweden.

Looking at the distributions of variable values, I observed that it is better to transform the GDP per capita and CO_2 emissions per capita variables to logarithmic form, as in that form it represents a more linear distribution (constant distance between non-empty bins of a histogram binning). Surprisingly, a better distribution is obtained also when taking the logarithm of oil rent percentages of GDP (countries have low oil rents, with "exponential" GDP). Therefore, these variables were transformed into logarithmic measures.

3 MODELS

There are three separate models used, in different settings:

1. Cross-sectional OLS: on data of 2005 and the latest available year (2023).
2. First-difference models incorporating time trends with varying lag structures (no lag, 2-year, and 6-year lags).
3. Fixed-effects models controlling for both country and time dummies.

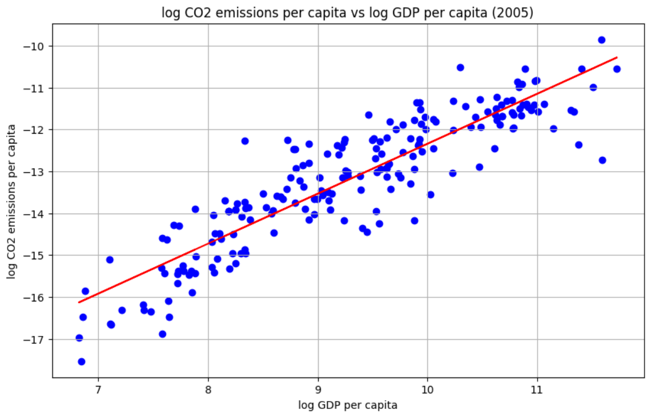
Model results are available in the table on the last page.

3.1 OLS (cross-sectional, 2005 and 2023)

Ordinary linear regressions, without accounting for any confounders, lead to biased results. The model fits estimated the impact coefficient at 1.192 and 1.06, respectively, which suggest strong positive and statistically very significant relationship with CO_2 emissions. There is a notable decrease in the coefficient for the second model, suggesting that the size of the impact may have decreased over time.

Including oil rents as a confounder reduces the GDP coefficient to 1.127. This decrease indicates that some of the initial relationship between GDP and CO_2 emissions was actually attributable to oil rents. The oil rent coefficient (0.03) is also significant, suggesting that oil rents themselves are positively associated with CO_2 emissions. Therefore, it is important to include it in the model.

Despite high (~80%) R2 values of fit (as seen on the image below), OLS models are biased as many confounders are not accounted for.



3.1 First difference

First difference models run OLS in a different setup that cancels time-invariant confounders, by examining differences. This is possible due to panel data, and in this case only confounders associated with changes in GDP and changes in CO_2 emissions matter.

In these models, measured coefficient is significantly reduced (e.g. 0.564) to about half of its size, indicating that the relationship between changes in GDP and changes in CO2 is weaker than the relationship between the levels. The time trend is not significant. The R-squared is low (8.4%), which is expected as the new model is expected to explain changes, which have more notable variance compared to levels.

Introducing (2-year) lags for GDP further reduces the current GDP coefficient (0.5316). Interestingly, the lagged GDP coefficients have mixed signs: The 2-year lag is positive, indicating that the impact of GDP changes on CO_2 changes might take some time to materialize. With 6-year lags, the coefficient of the 6-year lag is statistically significant, unlike other lag coefficients, implying that impact may happen on a long-term (6+ years) on GDP changes on CO2 changes, and may also highlight more-than-6-year lag impacts.

Adding the change in oil rents as a confounder has almost no effect on the GDP coefficient. The change in oil rents is only marginally significant, suggesting that the change in oil rents is not strongly associated with the change in CO2. This also showcases the power of the method, the ability to control for time-invariant confounders.

Notably, in the 6-year lag model GDP coefficient increases (0.577) relative to the 2-year lag model. This is due to data exclusion (years 1992-1997 are dropped due to the 6-year lag), and when inspected more carefully, the effect vanishes, showing the importance of taking account of changes in results due to using different data(table).

First Difference, 2-Year Lags, with Oil Rents:

These models show a much weaker relationship between changes in GDP and changes in CO2 compared to the levels. The effect of the change in oil rents is small.

3.2 Fixed effect

These models model with time and country fixed effects, with parameters for each country and year. They can control for all time-invariant variables including country-specific effects, and can also control for global time trends.

In this model (without oil rents), the GDP coefficient (0.657) is higher than in the first-difference models, suggesting that country-specific unobserved factors are important, and impact is higher than previously estimated. Within R2 is measurable for this model, which is significantly high (0.3126), showing good prediction of variation within countries over time.

Adding oil rents has a minimal effect on the GDP coefficient when using the same data. The R-squared increases slightly.

These models suggest that unobserved country-specific factors play a substantial role in explaining CO2 emissions, increasing the impact coefficient to 0.657.

3.4 General findings

The coefficient of X is significant across all models but decreases in magnitude when additional controls and fixed effects are included. Oil rents have a significant effect on CO_2 emissions, reinforcing its role as a confounder, however fixed effect and first difference models can account for it, and change in oil rents is not significant. The inclusion of country and time dummies helps control for unobserved heterogeneity. Also, excluding observations due to missing oil rent data shifts coefficient estimates.

3.5 MODEL RESULTS

The table below describes the measures of all model fits.

Model	OLS 2005	OLS 2023	OLS 2005 with ***z*** (oil rent)	FirstDiff time trend, no lags	FirstDiff time trend, 2Y lags	FirstDiff time trend, 2Y lags, with Δ**z**
$\boldsymbol{\beta_1}$	1.192	1.063	1.134 (-0.06)	0.5635	0.5316	0.567 (vs. 0.566) (-0.001)
$\boldsymbol{\beta_z}$	-	-	0.030	-	-	-0.0007
p-val $\boldsymbol{\beta_z}$	-	-	~0	-	-	0.240
other $\boldsymbol{\beta}$	-	-	-	~0 (Δ time trend)	~0 (Δ time trend)	~0 (Δ time trend)
$\boldsymbol{R^2}$	0.809	0.814	0.850 (+0.41)	0.084	0.077	0.082 (vs. 0.082), (+0)

Model	FirstDiff time trend, 6Y lags	Fixed effects: time and country FEs	Fixed effects: time and country FEs, with ***z***
$\boldsymbol{\beta_1}$	0.577	0.658	0.665 (-0.001)
$\boldsymbol{\beta_z}$	-	-	-0.0011
p-val $\boldsymbol{\beta_z}$	-	-	~0.172
other $\boldsymbol{\beta}$	~0 (Δ time trend)	-	-
$\boldsymbol{R^2}$	0.091	0.176	0.177 (+0.0003)
within $\boldsymbol{R^2}$		0.313	0.322 (+0.0005)

Values inside parentheses describe the value gathered when running the analysis on the same, restricted data (excluding instances in 2022 and 2023), and the difference between the value and the measured result.

4. SUMMARY

The models reveal high significance in CO_2 emissions the GDP per capita values, with the best model suggesting a 0.657 increase in CO_2 emissions for every unit increase in GDP.

5. REFERENCES

[1] The World Bank, ‘World development indicators’, 2012, [Online]. Available: https://data.worldbank.org/data-catalog/world-development-indicators