Assignment 2 DA 4: Panel Practice

Anton Shestakov

February 2025

Intro & Data Description

Code and Data: https://github.com/Anton21a/DA4_homework.git

Research topic: CO2 emission and GDP

Research question: To what extent does economic activity cause CO2 emission?

The employed data was collected within *Our World in Data*, non-profit project affiliated with the University of Oxford. The total number of observation is 5146 including 166 countries observed between 1992 and 2022.

List of variables:

- country - name of country
- code - ISO
- year - years
- CO2_capita - per capita annual CO2 emissions in tonnes
- GDP_capita - GDP per capita in constant prices $, 2011
- population - number of citizens
- urban_pop - number of citizens living in cities
- coal_capita - annual coal consumption per capita in kWh
- oil_capita - annual oil consumption per capita in kWh
- gaz_cap - annual gas consumption per capita in kWh

Dependent variable is defined as CO2_capita, independent variable is GDP_capita. Both represent the indices per capita: annual emission CO2 (in tonnes) and gross domestic product measured in PPP usd at constant prices (2011) for a particular country, respectively.

As part of the data modification process, the independent variable CO2_capita was transformed due to its values being less than one, which would result in negative values after applying a logarithm. Using logarithmic transformation is a crucial step in the analysis, as it helps smooth the right-skewed distribution of the data, as illustrated in Figure 1 [Appendix]. To preserve as many observations as possible, 1 value was added to each CO2_capita value, ensuring that all values remain positive before applying the logarithm.

The study employs gas, oil and coal consumption within countries per capitaita, and share of urban population as potential confounders. The descriptive statistics is shown in Appendix [Table 5 & Table 6]. The variables were incorporated in some regressions due to their prospective impact on both the dependent and independent variables. The study also demonstrates a correlation matrix with quantitative results of potential relationships between incorporated variables [Table 7, Appendix].

The percentage of citizens who lives in cities might explicitly effect the life quality that is associated with different facilities based on CO2 emissions (cars, using fabric items, e.t.c). Along with that, urbanization might be one of key factors for productivity growth that might lead to positive changes in GDP per capita.

Gas, oil, and coal consumption per capita are also strong confounders in a model analyzing the relationship between GDP per capita and CO2 emissions per cap because they influence both the independent and dependent variables. Higher GDP per capita is typically associated with greater industrial activity, transportation and electricity demand, leading to increased fossil fuel consumption.

At the same time, fossil fuel consumption is a primary driver of CO2 emissions, as burning coal, oil, and gas releases greenhouse gases. If the fossil fuel consumption is not controlled, the estimated effect of GDP on CO2 emissions might be biased, as a significant portion of C02 emissions is directly tied to energy use rather than GDP itself. This could lead to an overestimation of GDP's impact, mistakenly attributing emissions growth to economic activity when, in reality, it is energy consumption driving the increase.

There are other potential confounders that could be included in the model, and the variables mentioned are just some of them. However, while the chosen model reduces the likelihood of endogeneity in the variation of the variable of interest, it may not eliminate it entirely. On the other hand, adding more control variables could introduce bias, especially when working with a limited sample size.

Models Estimation and Interpretation

Cross-section OLS for the year 2005 and for the last year

As it has been mentioned before, the data is exposed to right-skewness distribution of its values. Especially, it relates to the dependent (Y) and independent (X) variables. To smooth the distribution, the logarithmic transformation was undertaken to reduce the effect of outliers [Figure 2]

$$\log(\text{CO2\_capita}) = \beta_0 + \beta_1 \log(\text{GDP\_capita}) + \varepsilon \tag{1}$$

The regression results for both periods are presented in the table 1 below. Statistically significant coefficients are observed in the summary columns for 2005 and 2022. The findings indicate a positive relationship between the logarithm of GDP per capita and CO2 emissions per capita. Specifically, CO2 emissions per capita is associated with an increase by 0.715% and by 1.071% in 2005 and 2022, respectively, on average, with an increase of GDP per capita by 1% in the given periods.

Table 1: OLS Regression Models

	2005	2022	2005 incl. all
Constant	-5.051***	-9.215***	-4.993***
	(0.225)	(0.352)	(0.595)
Log of GDP per capita	0.715***	1.071***	0.595***
	(0.025)	(0.038)	(0.087)
Share of citizens in cities			-0.271
			(0.348)
Lof of Gas consumption per capita in kWh			0.153***
			(0.039)
Num.Obs.	163	163	69
R2	0.836	0.834	0.762
R2 Adj.	0.835	0.833	0.751
AIC	591.1	523.5	49.3
BIC	600.4	532.8	60.5
Log.Lik.	-74.861	-141.868	-19.660
F	817.908	809.400	69.542
RMSE	0.38	0.58	0.32
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001			

The third specification in the summary table depicts the results of regression that includes such confounders as share of citizents in cities and log values of gas consumption per capita. Other confounders are kept away this regression because of a small sample of 69 observations. Even though, added controls influence on coefficient of independent variable: CO2 emissions per capita in 2005 is associated with an increase by 0.595%, on average, ceteris paribus, with a 1% increase in GDP per capita in 2005. This represents a drop of more than 16% compared to the initial estimate.

First difference models with time trend

The second equation represents first difference model with time fix effect and no lags. The third equation is first difference model with time fix effect and two lags. Finally, the fourth equation is first difference model with time fix effect and six lags.

$$\Delta \ln(\text{CO2\_capita})_{it} = \beta_1 \Delta \ln(\text{GDP\_capita})_{it} + \gamma_t + \varepsilon_{it} \tag{2}$$

$$\Delta \ln(\text{CO2\_capita})_{it} = \sum_{k=0}^{2} \beta_k \Delta \ln(\text{GDP\_capita})_{i,t-k} + \gamma_t + \varepsilon_{it} \tag{3}$$

$$\Delta \ln(\text{CO2\_capita})_{it} = \sum_{k=0}^{6} \beta_k \Delta \ln(\text{GDP\_capita})_{i,t-k} + \gamma_t + \varepsilon_{it} \tag{4}$$

The regression results for each specification are presented in Table 2. As observed, the coefficient of the independent variable in period t adjusts as additional lags are introduced. Meanwhile, the first difference of the logarithmic value of GDP per capita in period t remains highly significant across all three specifications but declines in magnitude when considering for past GDP effects.

In the first specification, which includes a time trend but no lags, on average, a 1% increase in GDP per capita in period t tends to be followed by 0.675% increase in CO2 emissions per capita at the same. Given the first-difference transformation, this result suggests that accelerating GDP growth is linked to an acceleration in CO2 emissions growth, rather than changes in their absolute levels.

The second specification includes 2 lags that adjust immediate effect in t period. With time trend, on average, a 1% increase in the first difference of GDP per capita in period t tends to be followed with a 0.276% increase in the first difference of CO2 emissions per capita at the same period.

To estimate long-run effects, the third specification is additionally run with time trend and 6 lags. With fixed time effects, on average, a 1% increase in

the first difference of GDP per capita in period t tends to be followed with a 0.303% increase in the first difference of CO2 emissions per capita in the same period. Besides finding this association, the results show that the FD of fourth lag significantly effect the dependent variable. The model demonstrates the highest value of within R squared of 7.6% across all three specifications, i.e the built regression explains 7.6% variation in Y within each unit over time.

Table 2: Fixed Effects Regression Models in First Differences

	no lags with time FE	2 lags with time FE	6 lags with time FE
FD log of GDP in t	0.675***	0.276***	0.303***
	(0.028)	(0.032)	(0.038)
FD log of GDP in t+1		0.054	-0.012
		(0.037)	(0.021)
FD log of GDP in t+2		-0.001	0.013
		(0.005)	(0.038)
FD log of GDP in t+3			-0.070
			(0.048)
FD log of GDP in t+4			0.060*
			(0.027)
FD log of GDP in t+5			0.002
			(0.022)
FD log of GDP in t+6			-0.001
			(0.003)
Num.Obs.	5052	4726	4074
R2	0.676	0.121	0.146
R2 Adj.	0.674	0.115	0.139
R2 Within	0.674	0.068	0.076
R2 Within Adj.	0.674	0.067	0.074
AIC	-5998.3	-13257.6	-12210.5
BIC	-5789.4	-13050.8	-12008.5
RMSE	0.13	0.06	0.05
Std.Errors	Newey-West (L=24)	Newey-West (L=24)	Newey-West (L=24)
FE: year	X	X	X
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001			

Fixed effects model with time and country fixed effects

The difference between the fifth and sixth equations comes down to embedding confounders to the latter one. Both regressions takes into account time and country fixed effects. The former implies aggregate trend which are time-dependent pattern that affect all units in the dataset in the same way over time. Thus it helps to capture global time changes. The latter is denoted by delta coefficient to capture all time-invariant characteristics of country.

$$\ln(\text{CO2\_capita})_{it} = \beta_1 \ln(\text{GDP\_capita})_{it} + \gamma_t + \delta_i + \varepsilon_{it} \tag{5}$$

$$\begin{aligned}\ln(\text{CO2\_capita})_{it} = \beta_1 \ln(\text{GDP\_capita})_{it} + \beta_2 \text{share\_urban}_{it} + \beta_3 \ln(\text{coal\_capita})_{it} \\ + \beta_4 \ln(\text{oil\_capita})_{it} + \beta_5 \ln(\text{gaz\_capita})_{it} + \gamma_t + \delta_i + \varepsilon_{it}\end{aligned} \tag{6}$$

The summary Table 3 shows significant effect of the independent variable of interest on the dependent variable. Compared to its mean within cross-sectional unit, CO2 emission per capita (Y) is increased by 0.295%, on average, where and when GDP per capita (X) deviates from its unit-specific mean by 1%, conditional on aggregate trends in the world. The model exhibits a within R-squared value of 16.7%.

The second specification is controlled by embedded confounders which adjust the impact of logarithmic GDP per capita on the dependent value by almost two times. Thus, it reduces the risk of omitted variable bias by adjusting endogeneity in the variation of X.

Compared to its mean within cross-sectional unit, the dependent variable (Y) is associated with an increase by 0.152%, on average, ceteris paribus, where and when GDP per capita is higher by 1% compared to its mean within a cross-sectional unit, conditional on aggregate trends in the world. The model exhibits a high index of a within R-squared value of 64.8%.

To the same extent, confounders show significant estimates. Compared to its mean within a cross-sectional unit, the dependent variable (Y) is associated with an increase by 1.203%, 0.044%, 0.362%, and 0.041%, on average, ceteris paribus, and accounting for the time trend, when and where the share of urban citizens, the logarithmic value of coal consumption per capita, the logarithmic value of oil consumption per capita, and the logarithmic value of gas consumption per capita, respectively, are higher than their mean by 1% within a cross-sectional unit.

Table 3: regression models with time and country FE

	time&country FE	time&country FE + confounders
log of GDP per capita	0.295***	0.152***
	(0.042)	(0.036)
share of urban citizens		1.203***
		(0.263)
log of coal consumption per capita		0.044**
		(0.012)
log of oil consumption per capita		0.362***
		(0.032)
log of gas consumption per capita		0.041***
		(0.009)
Num.Obs.	5053	2121
R2	0.978	0.984
R2 Adj.	0.977	0.983
R2 Within	0.167	0.648
R2 Within Adj.	0.167	0.647
AIC	-5453.9	-4279.9
BIC	-4187.5	-3674.4
RMSE	0.14	0.08
Std.Errors	Newey-West (L=24)	Newey-West (L=24)
FE: year	X	X
FE: country	X	X
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

First difference model, with time trend, 2 year lags, with confounders

The panel regression in first differences includes the independent variable of interest with two lags, and also confounders on urban population and different non-renewable consumptions per capita.

$$\Delta \ln(\text{CO2\_capita})_{it} = \sum_{k=0}^{2} \beta_k \Delta \ln(\text{GDP\_capita})_{i,t-k} + \gamma \Delta \text{share\_urban}_{it} + \delta \Delta \ln(\text{gaz\_capita})_{it} + \theta \Delta \ln(\text{coal\_capita})_{it} + \lambda \Delta \ln(\text{oil\_capita})_{it} + \gamma_t + \varepsilon_{it} \tag{7}$$

The summary Table 4 shows significant estimate for immediate effect in period t

on the dependent variable. The estimate differs from the one obtained in the second specification in Table 2. That means the added confounders make a difference in estimation reducing the risk of OVB.

A 1% increase in GDP per capita in t period tends to be followed by a 0.263% increase in CO2 emissions per capita at the same period, conditional on aggregate trends in the world. The explained variation of Y within each unit over time is 21.6% which is almost three times higher than the same type of model but without the incorporated confounders.

Table 4: FD regression models with 2 lags and confounders

	2 lags with time FE
FD log of GDP in t	0.263***
	(0.060)
FD log of GDP in t+1	0.017
	(0.043)
FD log of GDP in t+2	-0.002
	(0.004)
FD Share of urban population in t	0.532*
	(0.217)
FD gas consumption per capita in t	0.019
	(0.011)
FD coal consumption per capita in t	0.018*
	(0.008)
FD oil consumption per capita in t	0.220***
	(0.025)
Num.Obs.	1963
R2	0.338
R2 Adj.	0.326
R2 Within	0.216
R2 Within Adj.	0.213
AIC	-6756.1
BIC	-6555.1
RMSE	0.04
Std.Errors	Newey-West (L=24)
FE: year	X
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

Appendix

Table 5: Descriptive Statistics

skim_variable	n_missing	complete_rate	numeric.mean	numeric.sd
country	0	1.0000000	NA	NA
code	0	1.0000000	NA	NA
year	0	1.0000000	2.007000e+03	8.945141e+00
C02_capita	93	0.9819277	4.764734e+00	6.556941e+00
GDP_capita	0	1.0000000	1.474046e+04	1.652082e+04
population	0	1.0000000	4.053467e+07	1.417606e+08
urban_pop	93	0.9819277	2.063552e+07	6.263688e+07
coal_capita	2124	0.5872522	4.319688e+03	6.526629e+03
oil_capita	2124	0.5872522	1.321359e+04	1.701750e+04
gaz_capita	2124	0.5872522	9.876907e+03	2.274847e+04

Table 6: Descriptive Statistics

skim_variable	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100
country	NA	NA	NA	NA	NA
code	NA	NA	NA	NA	NA
year	1992.000000	1.999000e+03	2.007000e+03	2.015000e+03	2.022000e+03
C02_capita	0.021731	6.386760e-01	2.588560e+00	6.695069e+00	7.662412e+01
GDP_capita	377.580080	3.059177e+03	8.808597e+03	2.125878e+04	1.600512e+05
population	66850.000000	3.716662e+06	9.918328e+06	2.746253e+07	1.426437e+09
urban_pop	31532.000000	1.795135e+06	4.606889e+06	1.459029e+07	8.975784e+08
coal_capita	0.000000	0.000000e+00	1.032000e+03	6.215500e+03	3.817800e+04
oil_capita	0.000000	2.046500e+03	8.847000e+03	1.844650e+04	1.411890e+05
gaz_capita	0.000000	6.100000e+01	4.344500e+03	1.052575e+04	2.874690e+05

Table 7: Correlation Matrix

	year	C02_cap	GDP_cap	population	urban_pop	coal_cap	oil_cap	gaz_cap
year	1.00	-0.04	0.25	0.03	0.07	-0.07	-0.01	0.03
C02_cap	-0.04	1.00	0.62	-0.11	-0.06	0.27	0.56	0.79
GDP_cap	0.25	0.62	1.00	-0.14	-0.08	0.19	0.70	0.47
population	0.03	-0.11	-0.14	1.00	0.94	0.09	-0.12	-0.08
urban_pop	0.07	-0.06	-0.08	0.94	1.00	0.16	-0.08	-0.06
coal_cap	-0.07	0.27	0.19	0.09	0.16	1.00	0.13	-0.04
oil_cap	-0.01	0.56	0.70	-0.12	-0.08	0.13	1.00	0.37
gaz_cap	0.03	0.79	0.47	-0.08	-0.06	-0.04	0.37	1.00

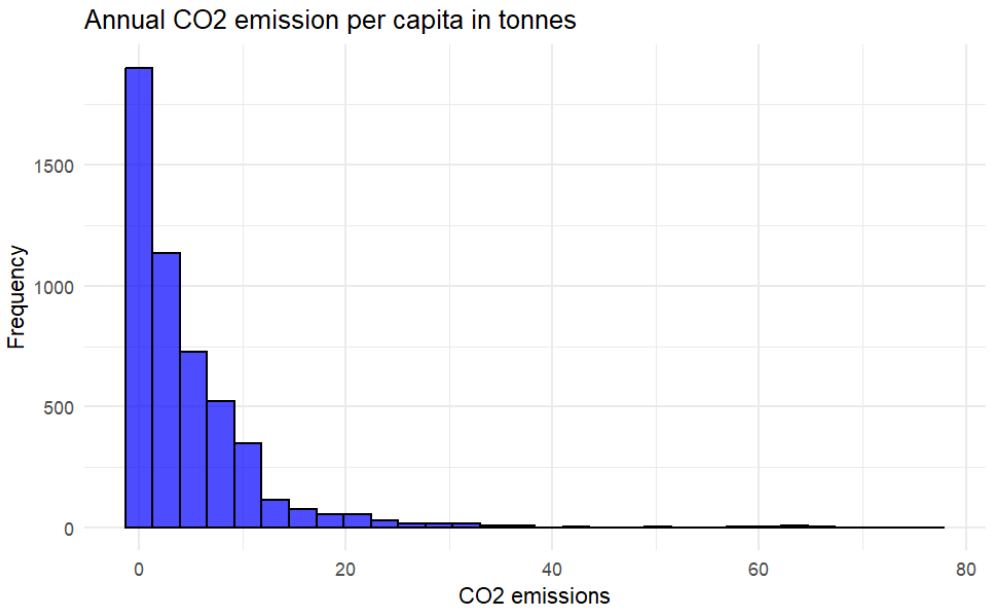

Figure 1: CO2 emission per capita in tonnes (annual)

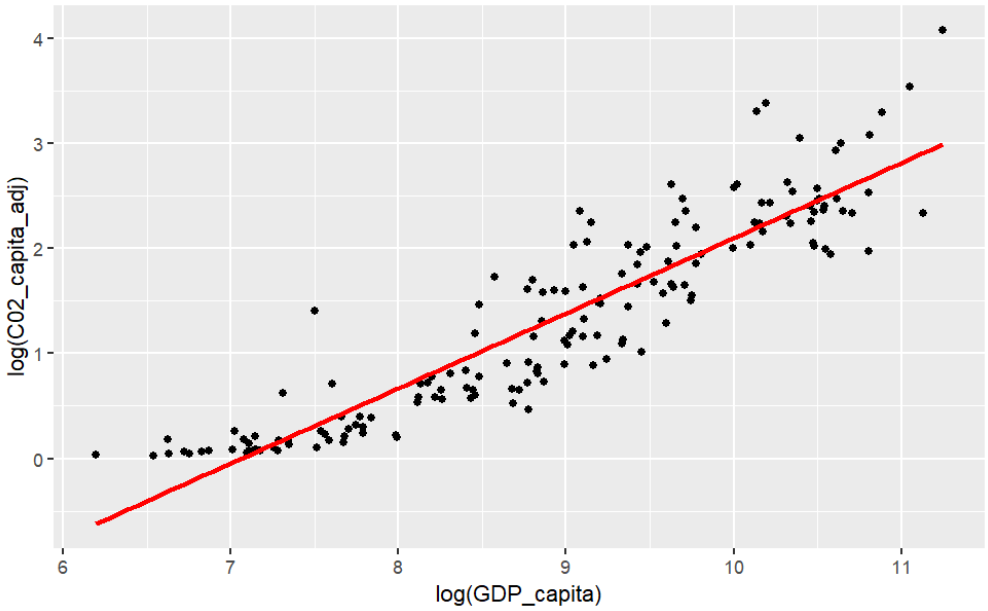

Figure 2: Log of GDP (in $) and log of CO2 (in tonnes) per capita for 2005