Author: Márton Nagy
Course: Data Analysis 4 – Causal Analysis (2nd Home Assignment)

Blame the wealthy? Examining the possible causal link between economic output and CO2 emissions[1]

1. Introduction to the source data

I have downloaded the indicators for the analysis from the World Bank's World Development Indicators website[2]. I have gathered the following variables. My outcome variable is *Carbon dioxide (CO2) emissions (total) excluding LULUCF[3] (Mt CO2e)*, which will be later transformed to a tons per capita measure (CO2 for short). My causal variable is *GDP per capita, PPP (constant 2021 international $)* (GDP p.c. for short). I have identified 4 possible confounder variables (I also indicate a possible mechanism for them). (1) Energy intensity of the economy (*Energy use (kg of oil equivalent) per $1,000 GDP (constant 2021 PPP),* energy intensity for short): as an economy gets more energy intense, it might have both a higher GDP and higher CO2 emissions. (2) Sectoral composition of the economy (*Industry (including construction), value added (% of GDP)*, industry share for short): industrialization of a country might entail higher GDP and higher emissions. (3) Renewable energy share in energy output (*Renewable electricity output (% of total electricity output)*, renewable share for short): higher use renewable sources might entail lower emissions, but higher GDP (as more developed economies turn towards renewables) (5) Urbanization (*Urban population (% of total population)*,): as more and more people move to the cities, emissions may get lower (as people need to commute much less), but GDP may grow (as people perform higher value-added jobs in large cities). I have also downloaded the *Population, total* variable to calculate per capita CO2 emissions (population for short).

As for the outcome, note that it only measures CO2 emissions, rather than total greenhouse gas (GHG) emissions, so it is an imperfect measure of GHG emissions. I have chosen this variable rather than total GHG emissions in CO2 equivalent units, as the task description explicitly asked for this. After transforming to a tidy long format, my raw data had 6944 country-year observations (217 countries observed through 32 years, from 1992 until 2023). The percentage of non-missing observations per variable are presented in Table 1.

Table 1: Percentage of non-missing country-year observations per variable

Variable name	Non-missing percentage
CO2	93.55
GDP p.c.	89.52
Energy intensity	45.97
Industry share	85.27
Renewable share	75.06
Urban population share	99.08
Population	100.00

2. Data cleaning

As I could do nothing with observations where either the outcome or the causal variables were missing, I dropped all such countries. This meant dropping 37[4] out of the 217 countries, leaving

[1] Source data and code are available at my public GitHub repository.
[2] World Bank (2025). *World Development Indicators*. https://databank.worldbank.org/source/world-development-indicators
[3] LULUCF means land use, land-use change and forestry. I have not found such a variable that would include LULUCF in CO2 emissions, thus I have gone with excluding this.
[4] The following countries were dropped (by country code): AFG, AND, ASM, BTN, CHI, CUB, CUW, CYM, DJI, ERI, FRO, GIB, GRL, GUM, IMN, LBN, LIE, MAF, MCO, MNE, MNP, NCL, PRK, PSE, PYF, SMR, SRB, SSD,

180 countries in my dataset (meaning 5770 country-year observations)[5].

Next, I looked at my possible confounders. As both energy intensity and renewable share still had a large proportion of missing values, I decided not to work with them further. Unfortunately, the remaining 7.6% missing values in industry share affected roughly a hundred of the countries still in my sample, so instead I decided to go on with the urban population share, which had no missing values at this point. After having filtered my sample, I added the CO2 per capita variable by simply dividing CO2 by population.

3. Descriptive statistics

Table 2 presents the summary statistics on my three main variables. We can see already that both GDP p.c. and CO2 p.c. have very skewed[6] distributions with extreme values. This indicates that it is indeed reasonable to log-transform these variables.

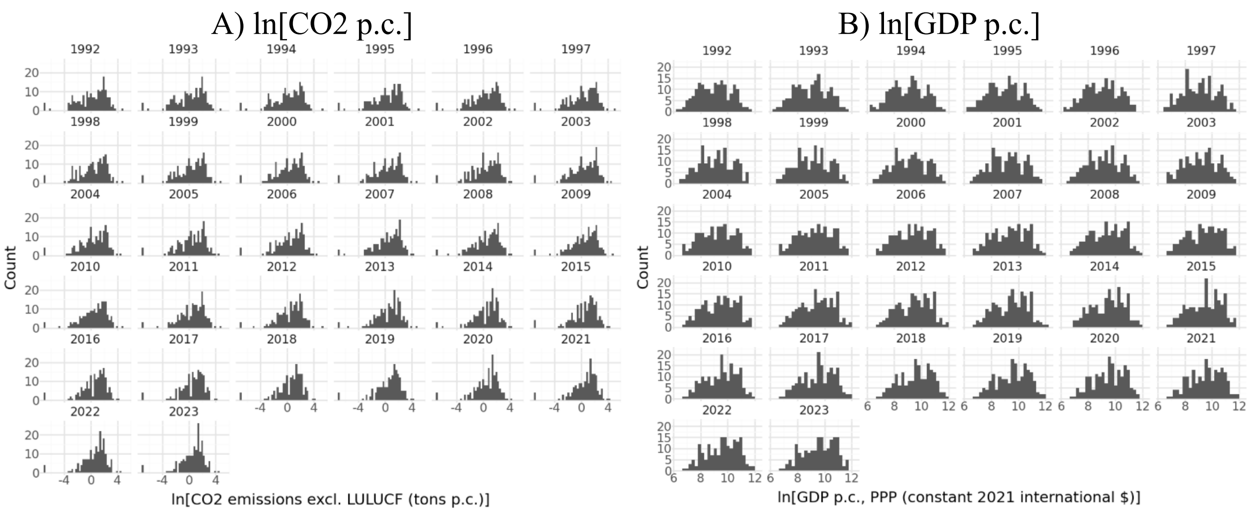
Table 2: Summary statistics of the untransformed variables

	Mean	SD	Min.	5%	10%	25%	50%	75%	90%	95%	Max.
CO2 p.c.	5.12	9.90	0.00	0.06	0.13	0.56	2.30	6.54	11.36	18.13	202.87
GDP p.c.	2.13e4	2.40e4	510.82	1492.49	2172.29	4160.62	1.24e4	3.11e4	5.43e4	6.65e4	1.74e5
Urban pop. %	56.17	23.44	6.29	18.38	23.59	36.53	56.21	74.64	84.42	94.16	100.00

Note that interestingly, the minimum of CO2 per capita is zero for some country-year observations. So, to take the logarithm of it, I have decided to adjust these values to a still minimal, but positive number[7]. This only affected 120 rows in my dataset (roughly 2% of the sample), so this should not substantially affect my results.

After taking the log of both the outcome and the causal variables, we can check the distributions graphically as well, broken down by years in the sample. This is presented in Figure 1, along with the distribution of the untransformed urban population share variable.

Figure 1: Distribution of key variables by year



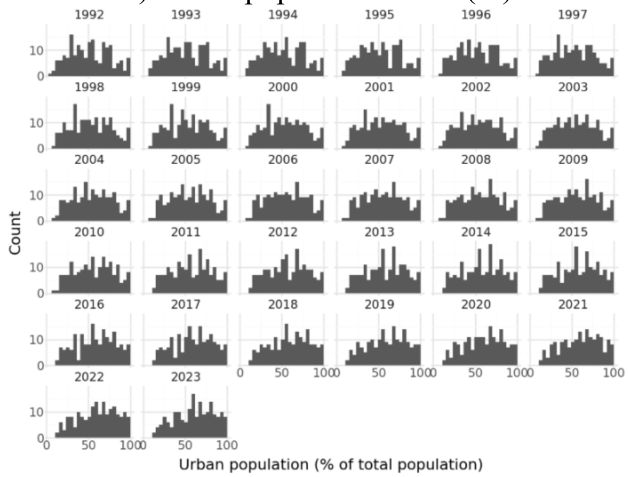
SXM, SYR, TCA, TON, VEN, VGB, VIR, XKX, YEM.

[5] Note that I have downloaded the data from the website with selecting all units that were classified as countries. This meant data for some autonomous territories as well – thus the larger number than the official 195.

[6] This is also shown by the histograms per year, omitted from this report because of spatial constraints. Please find the figures in the Jupyter notebook.

[7] More precisely, I changed zeros to the half of the minimum value (without zeros) before taking the logarithm.

C) Urban population share (%)



4. Model estimates

Table 3 presents the results of all estimated models. Note that I have decided to estimate all models weighted by population, as I believe these results are more informative, given that the models examine the link between per capita variables[8]. Also, from a policy point of view, it makes more sense to give more weight to larger, more influential countries in the estimates. The following models have been estimated:

- simple cross-sectional models for the years 2005 and 2023;
- first difference (FD) panel models with zero, two and six lags included;
- and a fixed-effects (FE) model.

The cross-sectional model for 2005, the FD with 2 lags and the FE have also been estimated by including the possible confounder, urban population share.

Table 3: Regression results (weighted by population)

	(1) OLS 2005	(2) OLS 2005 (control)	(3) OLS 2023	(4) FD	(5) FD 2 lags	(6) FD 2 lags (control)	(7) FD 6 lags	(8) FE	(9) FE (control)
Outcome variable	ln(CO2 p.c.)	ln(CO2 p.c.)	ln(CO2 p.c.)	Δln(CO2 p.c.)	Δln(CO2 p.c.)	Δln(CO2 p.c.)	Δln(CO2 p.c.)	ln(CO2 p.c.)	ln(CO2 p.c.)
ln(GDP p.c.)	1.14*** (0.09)	1.42*** (0.14)	1.27*** (0.11)					0.75*** (0.07)	0.69*** (0.09)
Δln(GDP p.c.)				0.63*** (0.04)	0.64*** (0.07)	0.61*** (0.07)	0.69*** (0.08)		
Δln(GDP p.c.) lag 1					-0.03 (0.05)	-0.06 (0.05)	-0.03 (0.05)		
Δln(GDP p.c.) lag 2					0.05 (0.04)	0.02 (0.04)	0.03 (0.04)		
Δln(GDP p.c.) lag 3							0.04 (0.03)		
Δln(GDP p.c.) lag 4							0.05 (0.04)		
Δln(GDP p.c.) lag 5							-0.02 (0.04)		

[8] I have also calculated unweighted models, the estimates of which can be found in the submitted Jupyter-notebook. Generally, those estimates suggest somewhat lower coefficients.

	(1) OLS 2005	(2) OLS 2005 (control)	(3) OLS 2023	(4) FD	(5) FD 2 lags	(6) FD 2 lags (control)	(7) FD 6 lags	(8) FE	(9) FE (control)
Outcome variable	ln(CO2 p.c.)	ln(CO2 p.c.)	ln(CO2 p.c.)	Δln(CO2 p.c.)	Δln(CO2 p.c.)	Δln(CO2 p.c.)	Δln(CO2 p.c.)	ln(CO2 p.c.)	ln(CO2 p.c.)
Δln(GDP p.c.) lag 6							-0.02 (0.05)		
Intercept	-9.50*** (0.98)	-11.24*** (1.07)	-11.20*** (0.99)	-0.00 (0.00)	-0.00 (0.00)	-0.01* (0.00)	-0.01** (0.00)		
Δln(GDP p.c.) cumul					0.66*** (0.05)	0.58*** (0.07)	0.74*** (0.05)		
Urban population %	no	yes**	no	no	no	yes	no	no	yes
Year dummies				yes	yes	yes	yes	yes	yes
Observations (T:N)	1:180	1:180	1:180	31:180	29:180	29:180	25:180	32:180	32:180
R^2	0.71	0.72	0.80	0.26	0.27	0.27	0.30	0.98	0.98
Within R^2								0.73	0.73
SE type	HC1	HC1	HC1	clustered	clustered	clustered	clustered	clustered	clustered

Notes: All models are weighted by population. Robust standard errors in parentheses (type indicated by the SE type *row). Cumulative effect estimates calculated by separate models but indicated under the respective FD model for simplicity. For FE models, simple R^2 value means the LSDV R^2, within R^2 is presented separately. Coefficients for the confounder variable are not shown, but statistical significance is denoted by stars. For the OLS and FE models, the confounder is added as level, while for the FD models, it is added as a first difference with appropriate lags. For models including lagged confounders, the stars relate to the most significant lag. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$*

Now, let's interpret the estimated coefficients. Note that for now, I will refrain from using causal language in the interpretations, as I will discuss in detail in Section 5 to what extent the results of certain models may be interpreted as a causal estimate. Also, I will not interpret the intercept estimates, nor all the lagged variable coefficients as these are rather uninformative on their own.

4.1. Cross-sectional models

Model (1) suggests that in 2005, countries with a 1% higher GDP p.c. had on average 1.14% higher CO2 emissions p.c. According to Model (2), comparing two countries with the same urban population share, the one with a 1% higher GDP p.c. had on average 1.42% higher CO2 emissions p.c. in 2005. This suggests a slight omitted variable bias[9] when not controlling for urban population share. Model (3) uncovers that the pattern of association between our causal and outcome variables is relatively stable over time: in 2023, countries with a 1% higher GDP p.c. had on average 1.27% higher CO2 emissions p.c.

4.2. FD models

Model (4) tells us that comparing two countries for the same year or comparing different years for the same country, controlling for the aggregate trend in a flexible way, CO2 emissions p.c. tends to change by 0.63% for observations where GDP p.c. changes by 1%[10]. Adding two lags of the causal

[9] Note that in this cross-sectional sample, the bias is actually in the opposite direction then what I would have expected given the hypothesized mechanism in Section 1. However, as we will see later, this is only this way in the cross-sectional set-up.

[10] Note that the coefficients relate to a 1 log unit change in GDP p.c., thus a 0.01 log unit change relates to a 1% change.

variable to the model in Model (5) does not practically make a difference in the contemporaneous coefficient (0.64% change associated with a 1% change in GDP p.c.). The coefficients on the lagged variables are rather small and statistically insignificant. The cumulative estimate uncovers that within 2 years of a 1% GDP p.c. change, CO2 emissions p.c. change by 0.66% (that is practically the same as the contemporaneous slope on its own, meaning that there is hardly any lagged relationship.

Adding the confounder urban population share in Model (6) results in a marginally smaller contemporaneous and cumulative estimate, suggesting a slight omitted variable bias. Note, however, that in this set-up, neither the contemporaneous nor the lagged confounders were statistically significant.

Model (7) suggests that comparing two countries for the same year or comparing different years for the same country, controlling for the aggregate trend in a flexible way, CO2 emissions p.c. tends to change by 0.69% for observations where GDP p.c. changes by 1%, when there was no change in GDP p.c. in the preceding 6 years. The lagged coefficients are again small and statistically insignificant. The cumulative estimate tells us that within 6 years of a 1% GDP p.c. change, CO2 emissions p.c. change by 0.74%

4.3. FE models

Model (8) shows that CO2 emissions p.c. is larger by 0.75% on average compared to its mean within countries and its mean within the year, where and when GDP p.c. is higher by 1% compared to its mean within countries and its mean within the year. Note that this is practically the same as the cumulative estimate in Model (7) which shows that FE models indeed tend to estimate long-run relationships. Adding urban population share to the model in Model (9) changes the estimate to 0.69% which is still in the 95% confidence interval of the previously mentioned estimate, but provides some evidence for a slight omitted variable bias in the initial model.

5. Causal interpretation possibilities

Generally, I believe that neither of the simple cross-sectional models can uncover a causal relationship, as it would be hard to believe that GDP p.c. is truly exogenous in these models, even if we control for the urban population share. There might be many more confounders in play (e.g. those listed in Section 1), the exclusion of which introduces a bias in the results.

As for the FD models, the extent to which they may be interpreted as a causal relationship depends primarily on whether the parallel trends assumption is satisfied. To assess this, I re-estimated Model (6) including two lead terms[11]. These results show that the one-period lead term is statistically significant at 5% (coefficient is 0.11), indicating that the parallel trends assumption may be violated. In this model, the cumulative estimate is 0.52, which is slightly lower than the previous 0.58 (but it falls within the 95% confidence interval of the previous estimate). So, the FD models are definitely closer to a causal interpretation than the simple cross-sectional models were, but as we have evidence that the parallel trends assumption may be violated, the estimates presented in Table 3 are rather only an upper bound of the causal effect. Also, even though FD models take care much of the endogenous variation in GDP p.c., we cannot be sure that there are no other relevant confounders not included in the models, correlated with changes in our outcome and causal variables (e.g., those listed in Section 1).

The FE models, especially Model (9), are also much closer to causality than the simple cross-sectional models were. Also, as the results of FE models coincide with the cumulative estimates of

[11] Detailed results may be found in the submitted Jupyter-notebook.

FD models, this shows that our long-term estimates are robust across these two model specifications. Importantly, however, the omitted variable bias may still be present in the FE models as well, just as mentioned in the case of the FD models.

Note that as outlined in Section 1, possible omitted confounders may be correlated with the outcome and causal variables in different ways. As a result, it is practically impossible to sign or assess the magnitude of the possible omitted variable bias in the models.

6. Summary and conclusion

In this paper, I have examined the possible causal link between GDP p.c. and CO2 emissions p.c. After estimating some simple cross-sectional models, I have also estimated FD and FE models to get closer to a causal interpretation. Even though I have included urban population share in some of my models, there still remains some uncertainty about the extent to which my results may be interpreted as true causal relationship. These stem from the fact that the parallel trends assumption may be violated (as suggested by a significant lead term), and also from a possible omitted variable bias, the direction of which cannot be plausibly signed. Nonetheless, my results suggest that a 1% increase in GDP p.c. may lead to around 0.6% increase in CO2 emissions p.c. in the same year, while the long-term effect is estimated to be around 0.7%. My results also shed some light on the fact that the relationship between GDP p.c. and CO2 emissions p.c. is mostly contemporaneous with little to no lagged effects.