ECON 412 Homework 3, due October 23rd 11:59pm

Assignment

October 23, 2023

Data Work: Greenhouse gas emissions and economic growth

During the class lecture on "Economic Growth and the Environment" by Grossman and Krueger (1995), we discussed the hypothesis of the environmental Kuznets curve. We observe an environmental Kuznets curve if improvements in environmental quality are correlated with GDP per capita in a particular way: at low incomes, increases in GDP per capita will tend to result in worsening environmental quality as the nations industrialize. At some point, incomes tend to rise enough that people seek improvements in environmental quality as they become richer, suggesting that the environmental Kuznets curve will have a downward-U shape in a graph of environmental quality versus GDP per capita.

In this assignment, you'll assess whether this relationship is valid for a different measure of environmental quality than the pollution measures discussed in class: greenhouse gas emissions.

You will work with data from a number of sources: population data from the United Nations World Population Prospects (UN WPP) (UN DESA, 2022); gross domestic product (GDP) from the Penn World Tables (PWT) (Feenstra et al., 2015); greenhouse gas emissions from coal, oil, natural gas, overall fuel consumption, and overall energy from the International Energy Agency (IEA) (IEA, 2023); and if you choose to explore more, territorial emissions, consumption emissions, and emissions transfers estimated by the Global Carbon Budget (GCB) project (Peters et al., 2012; Friedlingstein et al., 2022). The IEA and GCB estimates vary in part because of differences in estimation methods, and in part because the IEA only calculates emissions for fuel consumption whereas the GCB project estimates emissions throughout the supply chain.

The following folder contains raw data in the folder `data/01_raw`, as well as code for cleaning the raw data into something you can analyze and templates for analyzing and generating plots code in the folder `code`. First, copy these folders into a new folder, e.g. `C:/Projects/ECON-412` which I'll assume is the `home_folder` path for the duration of this assignment.[1]

It's advisable to set this new folder up with a sync through a cloud such as Dropbox, so that you have at least one cloud version (with version control so that you can access older versions of your code) and at least one local version.[2]

- https://u.pcloud.link/publink/show?code=kZJkXiVZF4odNFF025YC9AxgkMfhSLuVE8UV

- **Password:** `2023_Fall_ECON-412`

The Latex code which generated this document can be found at the following link (the same Template folder as prior assignments shared), in `ECON-412_hw/ECON-412_hw-03.tex`. You may find it useful to copy the contents of this file into your own Overleaf project and alter the copied Tex document for your own answers.

- https://www.overleaf.com/read/mpdhvnnjzsxq

The template code throughout this assignment performs a similar analysis as what you're being asked to do, but instead of looking at the relationship between emissions per capita and GDP per capita, it analyzes the relationship between life expectancy at various ages and GDP per capita. The bulk of your

---

[1]See the tutorial video `01_Downloading to a cloud and folder structure` on the Canvas media library if you're stuck.

[2]The rule of thumb typically involves three versions: a cloud, a local, and a back-up to an external hard drive. It's frustrating and time consuming to lose data or code because of technological snafus. Don't let it happen to you. In addition, your project team might find it useful to adapt the code from this assignment to your project, in which case having a cloud folder which can be shared will be useful.

coding task will be to adapt the code for the outcome variable of emissions per capita; you'll also be asked to review some of your econometrics and apply it to this scenario. Please note that none of the included code is certain to be the best, fastest, or easiest to understand, so please make improvements as you see fit. Before getting fancy, however, make sure you can get your code to *run*.

1 Data Cleaning

1. Adapt the R script in `code/00_master_run-of-show_template.R` to clean and merge all four datasets. To do this, the only thing you should need to change is the path of the local `home_folder` to the path that you've copied the `ECON-412` folder into.[3] Make sure the locals `clean_wpp`, `clean_pwt`, `clean_iea`, `clean_gcb`, and `merge_all` are set to 1 while the others are set to 0. That should be all you need to do to run the entire script. What's happening is the `00_master_run-of-show_template.R` is a "master" do file which calls upon a number of sub-files with the command `source()`. It's running the sub-file if the local associated with that file is set to 1, and ignoring the call to the sub-file if the local associated with that file is set to 0. Essentially, this "run of show" is toggling sub-files on and off.

   Each of the sub-files you've toggled on accomplishes a cleaning task, and the final sub-file merges them together. Setting up your cleaning first allows you to change your variable definitions, change what variables are included, or easily port one dataset to another project while having everything clearly organized and searchable. In general, try to do all your variable creation and adjusting in a cleaning file; otherwise it can be hard to find what variables you're using if they were defined deep in an analysis file.

   Once you've run this, the clean data should appear in your `data/03_clean` folder.[4]

2. Explore the data a little bit: what are the earliest and latest years available for the WPP, PWT, IEA, and GCB? Is there a difference in data availability across the IEA and GCB?

3. Describe how the availability of data for different years or countries might affect your interpretation of the relationship between GDP per capita and emissions per capita. What observations are likely to be missing? What are some of the risks of making estimates without detailed underlying data?

4. What are the units of the population data, GDP data, IEA emissions data and GCB emissions data? (Hint: you may find it useful to consult the raw data files)

5. What are the units of the per-capita IEA emissions, per-capita GDP, and per-capita GCB emissions? (Hint: You will need to read through the cleaning scripts in order to determine this).

2 Conceptual Framework

We'll explore a few alternative specifications for our analysis.

2.1 Specifications

Equation 1 describes a cross-section specification, showing GHGpc, greenhouse gas emissions per capita in country i and during a particular year t, as a function of per-capita GDP called GDPpc. Equation 2 instead takes $\log(\text{GHGpc})$ as the outcome variable and regresses on $\log(\text{GDPpc})$, while Equation 3 shows a regression of $\log(\text{GHGpc})$ on GDPpc.

---

[3](e.g.   set `home_folder <- file.path(C:,Projects,ECON-412)`.   You'll need to do this in both `00_startup_master.R` and `00_master_run_of_show.R`. Using `file.path()` is better than manually writing the path out because Mac and PC have different conventions about whether the forward or backward slash demarcate folders. Depending on your time of downloading, it's possible that you may have to check this within the sub-scripts, e.g $iea_clean.R$.

[4]The script `01_download_datasets_in_use.R` provides example code for downloading the data from R rather than manually clicking to the websites. It's good practice for replicability to have the data download process also be part of your code if your data are public. Since the IEA data requires a login in order to download and it's not strictly necessary to have R do the downloading for you, we'll skip that part.

$$\text{GHGpc}_{i,t} = \beta_0 + \beta_1 \text{GDPpc}_{i,t} + \varepsilon_{i,t} \tag{1}$$

$$\log(\text{GHGpc})_{i,t} = \beta_0 + \beta_1 \log(\text{GDPpc})_{i,t} + \varepsilon_{i,t} \tag{2}$$

$$\log(\text{GHGpc})_{i,t} = \beta_0 + \beta_1 \text{GDPpc}_{i,t} + \varepsilon_{i,t} \tag{3}$$

Equation 4 adds in a quadratic and a cubic term for GDP per capita, still within a particular year t. Equation 5 adds country-level fixed effects γ_i and time-varying fixed effects ν_t, implying the use of a panel analysis (i.e. the regression would be run over multiple years, rather than just a single year).

$$\text{GHGpc}_{i,t} = \beta_0 + \beta_1 \text{GDPpc}_{i,t} + \beta_2 \text{GDPpc}_{i,t}^2 + \beta_3 \text{GDPpc}_{i,t}^3 + \varepsilon_{i,t} \tag{4}$$

$$\text{GHGpc}_{i,t} = \beta_0 + \beta_1 \text{GDPpc}_{i,t} + \beta_2 \text{GDPpc}_{i,t}^2 + \beta_3 \text{GDPpc}_{i,t}^3 + \gamma_i + \nu_t + \varepsilon_{i,t} \tag{5}$$

1. For Equations 1 to Equation 3, carefully explain the interpretation of the β coefficients in a way that someone not familiar with regressions could follow. [5]

2. Describe what effect the inclusion of country- and year-level fixed effects has in Equation 5 on the interpretation of this particular regression. What is the difference in interpretation between the β_1 in Equation 4 and 5? In general, what is the difference in interpretation of the coefficients in a cross-section versus a panel data which contains both country and year fixed effects?

3. What would be a little odd if we were to run Equation 4 as a panel over multiple years? [6]

4. Explain when we may want to include country and time fixed effects, and when we might *not* want to include them. Why in this case does it make sense to include both, rather than one or the other?

5. Explain why it may be logical to include the cubic and quadratic terms in Equation 4 and Equation 5. Provide another example of a real-world regression in which quadratic or higher polynomial powers could make intuitive sense.

3 Analysis

3.1 Summary Statistics

1. Save the code in `code/03_analysis/03_analysis_hw-03_template.R` as a new file, e.g. `03_analysis_hw-03.R`. Adapt this code to generate a table of the summary statistics for the regressions given by Equations 1 to 5. Your table may look similar to Table 1. Include two different measures for greenhouse gases per capita. Indicate which is your preferred measure that you'll use throughout the remaining analysis. Describe your reasoning behind choosing this particular measure (I'm interested in your reasoning process, not your exact answer). Make sure your table notes provide units and cite your data sources.

---

[5]Hint: For the regressions involving logarithms, how does this interpretation relate to elasticities or semi-elasticities, i.e. percentage changes? If you need a refresher on your econometrics interpretations, a good resource is Ben Lambert's Youtube channel here.

[6]Note that the same issue were to arise if you ran Equation 1, Equation 2 or Equation 3 as a panel over multiple years. This strategy would be called a **pooled regression**.

Table 1: Summary Statistics

Statistic	N	Min	Mean	Median	Max	St. Dev.
Life Expectancy at birth	10,201	11.99	64.31	67.55	85.27	11.65
Life Expectancy, Age 15	10,201	8.69	55.23	56.17	70.48	6.50
Life Expectancy, Age 65	10,201	3.74	14.13	13.71	22.59	2.46
Population (insert units)	10,201	4.26	31,314.14	6,367.52	1,419,730.00	115,703.20
GDP (insert units here)	10,201	58.00	310,727.50	31,628.73	20,860,506.00	1,225,322.00
GDP per capita (units)	10,201	240.70	13,150.80	6,463.14	283,927.30	19,058.55
GDP (units) per (units) capita	10,201	0.24	13.15	6.46	283.93	19.06

Notes: «Input your note about the units for GDP and its citation» Population and life expectancy data are from the United Nations World Population Prospects (UN DESA, 2022).

3.2   Exploratory Figures

1. Save the code in `code/04_plots/04_plots_hw-03_template.R` as `04_plots_hw-03.R`, and adapt this to generate scatter plots for the cross sections of GDP per capita on greenhouse gas emissions per capita during two years of your choosing (remembering to change the units, titles, and data sources), corresponding to Figure 1. Comment on and interpret these plots: do they accord with your intuition? Is there anything surprising or noteworthy about them?

2. Similarly adapt the code to generate the figure corresponding to Figure 2 but instead with of GDP per capita versus greenhouse gas emissions per capita for a particular country of your choosing. Comment on this figure: does it track with what you know about this country over time? Are there any patterns that you observe in this figure that you wouldn't have been able to observe in the cross section? Similarly, are there patterns that show up in the cross section that you wouldn't have been able to observe looking at just a single time series?
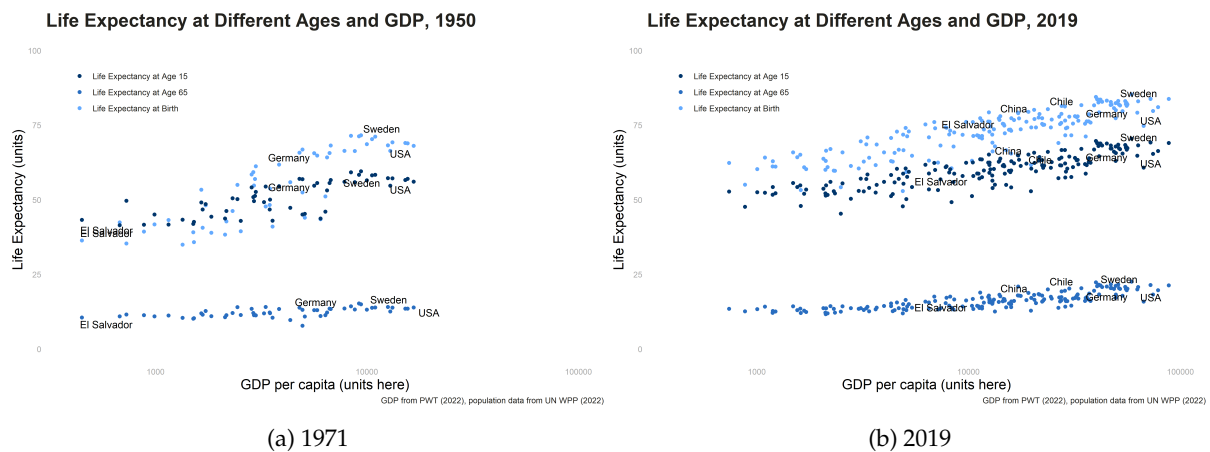

(a) 1971

(b) 2019

Figure 1: Comparison of Life Expectancies to GDP per capita over time
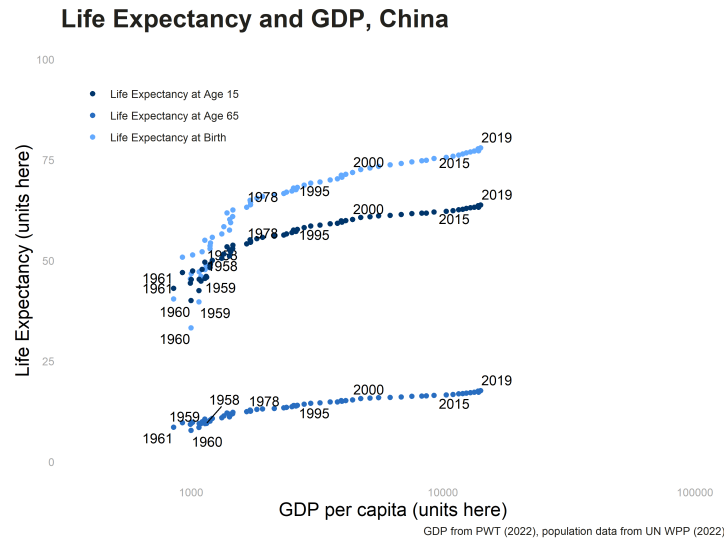
Life Expectancy and GDP, China

Figure 2: China's Evolution, 1950-2019

3.3 Results

Now let's run some regressions and generate some predictions.

1. Adapt your code in `code/03_analysis/03_analysis_hw-03.R` to generate two tables that correspond to Equations 1 through 5 above. One of these tables should include the results of running regressions from Equations 1 through 3 on the same two years you showed for your figures in Figure 1. The code will be similar to that which generated Table 2.

2. The other table should include the results from Equation 4 and Equation 5, with the regression for Equation 4 restricted to the most recent year, and Equation 5 a panel over multiple available years. You may find it useful to adapt the code which produced Table 3.[7]

3. Interpret the results for each column of each table in words (no need to be repetitive, but make clear the distinctions between them). Describe the economic and statistical significance of these coefficients. What is your preferred specification and why? Are any of these results surprising or otherwise noteworthy?

4. Adapt the code of `code/04_plots/04_plots_hw-03_template.R` to generate a scatter plot with GDP per capita on the X axis, and predicted values for greenhouse gas emissions per capita on the Y axis, according to the specification of Equation 5. This corresponds fairly closely to the exercise of Grossman and Krueger (1995), yielding a plot like Figure 3. Do you observe an environmental Kuznets curve? Comment on why or why not. Has your analysis made a causal argument? Comment on why or why not.

---

[7]It's your choice whether you'd like to restrict the data to a balanced panel, i.e. where all years and all units are available, or whether you'd like to use an unbalanced panel in which you include whatever country-year combinations are available. But whichever you do, spend a couple minutes to think about what the trade-offs are.
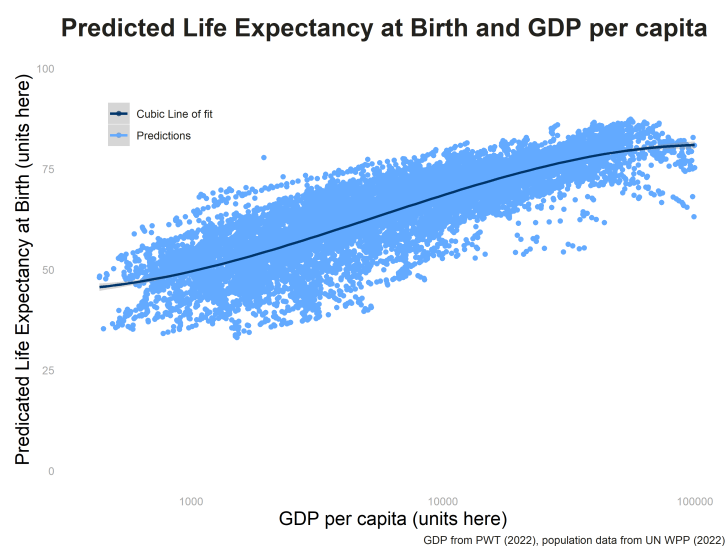
Figure 3: Predicted Values

Table 2: Life Expectancy Cross-Section Comparisons, 1950 and 2019

	Birth		Age 15		Age 65	
	1950	2019	1950	2019	1950	2019
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	42.823	67.733	42.617	42.102	11.127	14.266
	(1.676)	(0.665)	(6.361)	(4.614)	(0.224)	(0.201)
GDP per capita	0.002	0.000				
	(0.000)	(0.000)				
Log GDP per capita			2.073	2.359		
			(0.651)	(0.401)		
TFR			-1.880	-1.671		
			(0.288)	(0.363)		
GDP (unit) per (unit)					0.213	0.093
					(0.028)	(0.009)
Num.Obs.	55	182	55	182	55	182
Mean	54.870	73.080	50.660	60.220	12.260	16.370
R2	0.611	0.508	0.769	0.724	0.360	0.572
R2 Adj.	0.604	0.505	0.760	0.721	0.348	0.569
F	78.666	98.508	88.414	209.054	56.186	108.220

Notes:

Robust standard errors given in parentheses. Population and life expectancy are obtained from UN DESA (2022). Gross domestic product (GDP) is (what units) from Feenstra et al. (2015).

Table 3: Panel Regression, Life Expectancy

	Life Expectancy at Birth			
	1950	2019	Within	Twoway
	(1)	(2)	(3)	(4)
Intercept	33.437	62.576		
	(2.405)	(0.789)		
GDP (units) per (units) capita	6.714	0.891	0.821	-0.141
	(1.609)	(0.082)	(0.065)	(0.048)
(GDP(units) per (units))^2	-0.373	-0.014	-0.008	0.001
	(0.217)	(0.002)	(0.001)	(0.000)
(GDP (units) per (units))^3	0.006	0.000	0.000	0.000
	(0.008)	(0.000)	(0.000)	(0.000)
Mean	54.87	73.08	64.31	64.31
Country FE	N	N	Y	Y
Time FE	N	N	N	Y
Num.Obs.	55	182	10201	10201
R2	0.743	0.694	0.785	0.929
R2 Within			0.302	0.035

Notes:

Heteroskedascicity-robust standard errors clustered at the country (for within) and country and year (for twoways) levels given in parentheses. Population and life expectancy data are from UN DESA (2022). Gross domestic product (GDP) is (what units) from Feenstra et al. (2015).

References

Feenstra, R. C., Inklaar, R., and Timmer, M. P. (2015). The Next Generation of the Penn World Table. *American Economic Review*, 105(10):3150–3182.

Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Bakker, D. C. E., Hauck, J., Le Quéré, C., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Anthoni, P., Bates, N. R., Becker, M., Bellouin, N., Bopp, L., Chau, T. T. T., Chevallier, F., Chini, L. P., Cronin, M., Currie, K. I., Decharme, B., Djeutchouang, L. M., Dou, X., Evans, W., Feely, R. A., Feng, L., Gasser, T., Gilfillan, D., Gkritzalis, T., Grassi, G., Gregor, L., Gruber, N., Gürses, Ö., Harris, I., Houghton, R. A., Hurtt, G. C., Iida, Y., Ilyina, T., Luijkx, I. T., Jain, A., Jones, S. D., Kato, E., Kennedy, D., Klein Goldewijk, K., Knauer, J., Korsbakken, J. I., Körtzinger, A., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lienert, S., Liu, J., Marland, G., McGuire, P. C., Melton, J. R., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S.-I., Niwa, Y., Ono, T., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Rosan, T. M., Schwinger, J., Schwingshackl, C., Séférian, R., Sutton, A. J., Sweeney, C., Tanhua, T., Tans, P. P., Tian, H., Tilbrook, B., Tubiello, F., van der Werf, G. R., Vuichard, N., Wada, C., Wanninkhof, R., Watson, A. J., Willis, D., Wiltshire, A. J., Yuan, W., Yue, C., Yue, X., Zaehle, S., and Zeng, J. (2022). Global Carbon Budget 2021. *Earth System Science Data*, 14(4):1917–2005.

Grossman, G. M. and Krueger, A. B. (1995). Economic Growth and the Environment. *The Quarterly Journal of Economics*, 110(2):353–377.

IEA (2023). Greenhouse Gas Emissions from Energy: Highlights.

Peters, G. P., Davis, S. J., and Andrew, R. (2012). A synthesis of carbon in international trade. *Biogeosciences*, 9(8):3247–3276.

UN DESA (2022). World Population Prospects 2022: Methodology of the United Nations Population Estimates and Projections.