Day 5: Policy Problem - A Practical Exercise

Instructors: Cristián Jara[1], Javiera Petersen[2] and Raimundo Smith[3]

The main objective of this exercise is put into practice some of the topics we have reviewed during the training course. Particularly, we focus on the evaluation of environmental policies and their relation with inequality. For this purpose, we study the following paper: *"Does Enviromental Policy Affect Income Inequality? Evidence from the Clean Air Act"*, AEA Papers and Proccedings, 2019 by Akshaya Jha, Peter H. Matthews. The abstract of the paper says the following:

> *This paper quantifies the impact of environmental policy on income inequality. We focus on the Clean Air act and the National Ambient Air Quality standards for fine particulate matter and ozone. Using a matched difference-in differences estimator, we find evidence that both standards increased inequality in market income and a measure of income that deducts per-capita air pollution damage from adjusted gross income. While pollution standards can reduce pollution levels and thus result in significant environmental benefits in aggregate, our findings suggest that these standards appear to distort the distribution of economic resources in complex, and at times unfortunate, ways.*

The exercise is divided into two parts:

1. Summary statistics of the main variables and indicators used in the study and visualize the evolution of them over the years.

2. Replicate main results of the paper.

Exercise I: Summary statistics at the state level

The goal of this exercise is to summarize the main variables and indicators used in the paper at the state level. For this purpose, the authors work with three data sets for the different outcomes they use in the paper. Particularly,*SOI_ zipcode_ data_ Update.dta* and *LUR_ PM25_ O3_ 1999_ 2015_ Block_ Group.csv.* Next we describe each data and the calculations you should with them.

Market income data

1. *SOI_ zipcode_ data_ Update.dta* disaggregation is at the county level, so the first thing we must do is group all the counties in their corresponding states. Create a variable using `egen` command that computes the mean of the adjusted gross income by state. Then group this variable with *state* and generate another variable using the `egen` command.

   * An option is to work with *temporary variables.* Write `help tempvar` in Stata for more information about it.

2. Use the variables previously created and keep in the data the states at the top of the distribution, at the bottom of the distribution and NY.

3. Merge by year this data with auxiliary inflation data named *inflation_ cpi.* Keep all the observations from the original data and those that do match from inflation data. Drop all the other observations.

[1] National Institute of Statistics Chile: cristianjara21@gmail.com
[2] University College London (UCL): javierapertersenm@gmail.com
[3] World Bank: raimundo.smith.m@gmail.com

4. Deflate income variable to 2011 US dollars.
5. Generate a variable named *gini* and compute within state-year gini.
6. Create a new data using the command `collapse` with the information of the inequality indicators. Specifically, compute the following within state-year inequality indicators:

   (i) Percentile 90

   (ii) Percentile 50

   (iii) Percentile 10

   (iv) Mean

   (v) Standard Deviation

   (vi) Gini

   (vii) IQR

7. Repeat this process for all years (2005-2015) in the data and then append the corresponding datasets.

   * An option is to work with *temporary files*. Write `help tempfile` in Stata for more information about it.

8. Visualize the evolution of the indicators over the whole period only for NY and the state at the bottom of the distribution. Use only percentile 90, percentile 50, percentile 10, gini and mean.

$PM_{2.5}$ DATA

1. Import *LUR_ PM25_ O3_ 1999_ 2015_ Block_ Group.csv* data and keep observations corresponding to year 2005 or higher. Also keep only those observations where the variable *pollutant* takes the value `pm25`.
2. Repeat steps 1,2, 5 and 6 from the market income data, but using the variable *pred_ wght* instead of gross income.
3. Repeat step 8 from the market income data

O_3 DATA

1. Import *LUR_ PM25_ O3_ 1999_ 2015_ Block_ Group.csv* data and keep observations corresponding to year 2005 or higher. Also keep only those observations where the variable *pollutant* takes the value `o3`.
2. Repeat steps 1,2, 5 and 6 from the market income data, but using the variable *pred_ wght* instead of gross income.
3. Repeat step 8 from the market income data

Exercise II: DID with a particular kind of matching

In this exercise we will replicate one of the results of the paper. Specifically, using panel variation in the stringency of environmental regulation generated by the National Ambient Air Quality Standards (NAAQS), authors find a negative effect on the $PM_{2.5}$ log gini from a policy changed introduced in 2006. In table 1 of Jha, Matthews, and Muller (2019), of the online appendix we observed that this effect is of -6,7%. Using these three data sets: *LUR_ PM25_ county.dta*, *county_ population_ long.dta* and *compliance_ history_ clean_ long.dta*, you are asked to compute the previously described result following these instructions:

Part 1: Setup

1. Open *LUR_PM25_county.dta* data, keep observations between years 2005 and 2015, drop missings from variable *fips*.
2. Using the variables *fips* and *year*, merge this data with population data named *county_population_long.dta* and keep only matched and unmatched observations from the original data.
3. Generate the Coefficient of variation (CV) and create variables corresponding to the logarithm of each one of the indicators.
4. Using the variables *fips* and *year*, merge this data with NAAQS Standards data and keep only the observations from the "master data".
5. Using the variables *fips* and *year*, merge this data with NAAQS Standards data and keep only the observations from the "master data".
6. Create a "compliance" indicator using each of the policy changes variables in the data (*indcom*). This variables should reflect the "Always attainment counties" and the "non attainment counties". Then save this data and named it *master_data*.[4]

Part 2: Compliance and non compliance files plus matching

1. For policy changes introduced for O3 in 2008 (*everindcom3*) and for $PM_{2.5}$ in 2006 (*everindcom10*) create two different data sets: One for each value of compliance indicator (0 or 1).
2. Each one of these new data sets should have: information only for year 2005 and for the corresponding value of each policy (e.g One data for O3 policy with non compliance and one data for compliance). Also, keep only*log gini* and *fips* variables and create a constant variable named *index* that takes the value 1.
3. Save each data with a distinctive name, because you will use them for the matching estimator.[5]

Matching

In this section, you will have to work with a particular matching estimator used in Jha, Matthews, and Muller (2019) and originally proposed by Heckman, Ichimura, and Todd (1997). For a completely understanding of the process, we strongly recommend to read appendix 4 from the online appendix of Jha, Matthews, and Muller (2019), that you can find in the folder "Materials" from day 5.

In sum, based on the 2005 concentration levels of $PM_{2.5}$ you will have to match each county that was ever out of attainment with the relevant standard between 2005–2015 to ten counties that were always in attainment with this standard during 2005–2015.

The complete process can be summarized as follows:

1. Using the command `joinby` and the variable *index*, merge the compliance and noncompliance files you just created.
2. Generate a new variable named *abs_diff*, wich is the distance between non-attainment and always attainment counties.[6]

[4] You could used a `tempfile`.

[5] To save some memory in your computer, we suggest to work with `tempfiles`. However, it is not mandatory.

[6] Use the function `abs()`.

3. Keep only those observations where always attainment counties are less than or equal to 10.
4. Keep non-attainment and always attainment counties variables and *abs_ diff*. Then save the data with a distinctive name.
5. Now open *master_ data* and separately, create two data sets: 1) Always attainment data (using the indicator variable created in part 1; 2) Non attainment data using the indicator variable created in part 1.

   ***Clue:** First save "Always attainment" data using `preserve` and `restore` commands, and then create "Non attainment data".

6. Using the command `joinby` and the variable *fips*, merge "Non attainment data" with the data you create on item 4 and save it with a distinctive name.

   * The variable *fips* should identify non-attainment counties in both data sets.

7. Open "Always attainment" data you created on item 5 and append it with the date you just saved on item 6. Then create a varable named *constant* that takes de value 1.
8. Sort the data by year and by always attainment counties and, create a variable named *weights* equal to the sum of the *constant* variable.

   ***Clue:** Use the command `egen` and the function `total`.

8. Replace *weight* by 1/*weight* and then replace the variable by 1 if it correspond to a ever non-attainment county.
9. Finally, using the command `egen` and the function `group()`, create a variable named *fips_ matched* that groups counties matched counties (ever non-attainment county an its 10 matched always attainment counties.

Part 3: DID regression and common trends graphs

Now you are ready to replicate Jha, Matthews, and Muller (2019) results.

DID regression

Following the notation used in the paper, the regression you must run is the following

$$log(Y_{i,t}) \quad = \quad \alpha_i + \gamma_t + \beta NA_{i,t} + \theta PREVNA_{i,t} + \varepsilon_{i,t} \qquad (1)$$

where $Y_{i,t}$ is the outcome of interest in county i in year t; α_i represents county fixed effects; and γ_t represents year fixed effects. $NA_{i,t}$ is an indicator that assumes a value of one if and only if county i is out of attainment with the relevant NAAQS standard in year t; $PREVNA_{i,t}$ is an indicator that assumes a value of one if and only if county i is out of attainment with the previous NAAQS standard for the same pollutant in year t. Finally, $\varepsilon_{i,t}$ is the error term.

For this purpose, you will have to:

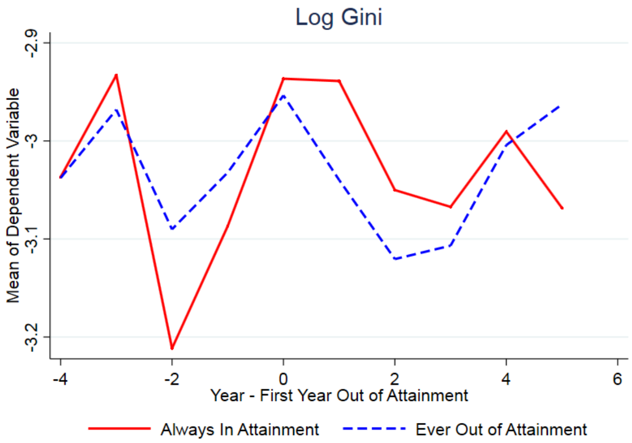
1. Use the command `reghdfe`, which is very useful when your specification have many fixed-effects.[7] Moreover, to replicate the result, you will have to use analytical weights, restrict the regression to years previous to 2015 and cluster errors by county.
2. Export the result using the command `outreg2`.

[7] See `help reghdfe` for more details on the command.

Common trends graphs

Finally, to replicate the commmon-trends assumption figure, you will have to graph the average outcome over time for ever non-attainment versus always-attainment counties. Certainly, there is not only one way make this graph, but figure 1 gives you and idea of how should it looks like:

Figure 1: Common Trends Assumption - Log Gini


Log Gini
Mean of Dependent Variable
Year - First Year Out of Attainment
Always In Attainment
Ever Out of Attainment