Data Assignment PPHA 45700

Introduction

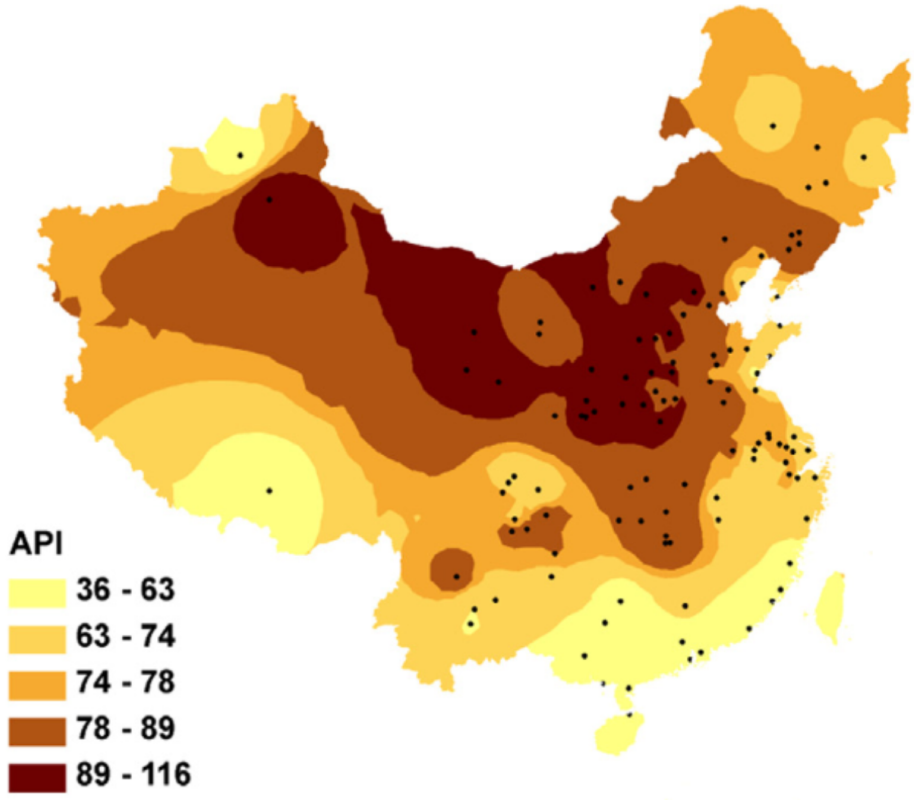
This assignment will take you through the process of examining real-world policy data and evaluating its effectiveness. It will require you to implement some of the different identification methods you have seen in the papers we have studied. I will create a folder in your name, put the input data files and your code into this folder, and then execute your code. You may use R or STATA but I strongly recommend R. The datasets are in R format, ask me if you need it in another format such as STATA.

Please also send an additional readme file in case there is information I should know to understand your output or code. For example if your code will generate a figure called "Figure1" and save it to the hard-drive, you can tell me to look for this. If your code will successively show figures on the screen let me know, so I will run your code bit by bit. If your code does *not* execute I will still do some basic debugging and corrections to fix any minor errors. But I will not correct any significant issues. Each question below has a specified number of points and most are "stand-alone" so you can miss one and still get the next right.

Your work must be original and done yourself. If you do not understand a question, come and talk to me and I will explain.

1. Perfection on demand

The Air Pollution Index is a measure of air quality used in China. A "blue-sky" day in China is defined as any day with an API below 100, a very generous interpretation of the term. The Figure below shows the API across China, averaged between 2001-2010. Large parts of the country have API values greater than or close to 100, indicative of a large number of days that are not blue-sky.



In order to motivate local officials to reduce pollution, the Chinese government introduced environmental compliance into the cadre promotion system for local city officials. Air quality measured by the Air Pollution Index (API) is the single most important indicator in the assessment. Cities were given points more points when they achieved a higher fraction of blue-sky days, with a perfect count (365 days in the year) being especially valued. Officials who obtained more points were more likely to be rewarded, and they were also ranked in a competitive process so that outperforming your peers had significant value.

A young PhD student discovers this policy and notes that the cut-off for whether a day qualifies as "blue-sky" or not is rather arbitrary. This seems a golden opportunity to observe officials who were rewarded, versus those who were not, and then write a paper on the effect of early promotions and employee recognition on job performance. The idea was to compare performance of officals who had many days where the API was just below 100 to those where many days had an API just above hundred. Once written, this seminal work would guarantee a position at the Harris School, with the opportunity to inflict courses with names such as "Environment and Development" on a captive audience.

The dataset China.Rdata provides the AQI for several years *for the city of Beijing* which is of naturally of particular interest. Plot the histogram of AQI for Beijing. Do you observe anything unusual? Might this affect the PhD student's planned research design, and if so why? (**5 points**)

2. Vehicle Rationing

A few years ago, the government of Delhi announced a series of emergency measures to control air pollution. This included a policy of rationing driving. The scheme worked as follows: Cars were classified into odd and even categories on the basis of the last digit of their license plates. It was mandated that only vehicles with odd numbered license plates could be driven on odd numbered dates and only those with even plates, on even dates. The restrictions were in force between the hours of 8 am and 8 pm for the first 15 days of January 2016. Recently they announced the policy would be repeated, starting November 4th. It is therefore of current interest to figure out whether this is a good idea.

I have uploaded data from the 2015 iteration of the policy to see if we can make sense of what happened last time around and whether pollution reduced. The very hard work of actually cleaning the pollution data we will use has been completed already.

Part 1: Read data and summarise variables

The dataset for the first part of the assignment is called "part1.Rdata". Read this into your statistical software. The definitions of variable names are as follows:

station_name = A string containing the name of a pollution monitoring station

station_id = A unique ID number for each site

PM10 = Particulate matter pollution reading (PM 10)

PM25 = Fine particulate matter pollution reading (PM 2.5)

NO2 = NO2 pollution reading

Hour = Runs from 1-24 and denotes the hour at which the reading was taken

Month = Runs from 1-12 and denotes the month

Day = Runs from 1-31 and denotes the day of month

time = A time variable that starts at 1 for Nov 01, 2015 and goes up by 1 each day

date = A full date variable

Delhi = A dummy / indicator variable that reads 0 when the station is outside Delhi

1. Output a summary of the variables PM25, PM10, NO2. In STATA you can do this using the 'univar' command. In R you can do this using the 'summary' command. The summary should provide minimum, mean, median, and maximum values. Also provide the number of missing values for each pollutant. (**1 point**)
2. Plot the average PM 2.5 reading over time i.e x-axis should be a continuous variable running from November 1 to April 29 and the y-axis should be the daily PM25 reading (averaged over all stations). Fit a smoothing curve through the points. Label x and y axes and provide a title. In R, one way to make a plot with a smoothing curve is to use the 'scatter.smooth' command (**2 points**)
3. Calculate the PM 2.5 levels for every hour of the day, averaging over all the days and stations in your dataset. Draw a line plot of pollution (y-axis) against hour-of-day (x-axis). Comment on what you see - are there particular times of day when pollution is really bad? Do these times line up with the hours that the vehicle rationing is in force? (**2 points**)
4. Find the average PM25 reading for each station **in Delhi** over the entire period of time for which you have data. Make a bar graph showing this average for each station. To focus on Delhi stations you need to drop all observations where the variable *Delhi* is 0. Can you think of reasons why some stations might be more polluted than others? (**2 points**)
5. Using the computation in (4), assume that Delhi began installing one pollution monitor each year starting from the worst polluted areas and then in successively less polluted areas, where the pollutant in question is PM 2.5. The city government reports average pollution as being the PM 2.5 reading averaged over *the stations that exist*. Calculate the annual pollution that would be reported by the Delhi government year on year starting when only one monitor existed, then two, then three, and so on. Plot these numbers on a figure (line plot). What is the reported change in pollution when all the monitors are installed versus when only the first was installed. (**1 points**)

Part 2: Evaluate the performance of vehicle rationing

I've made a separate Part 2 dataset called 'part2.Rdata' which has the following variables:

time = Time variable that starts at 1 for November 01, 2015 and goes up by 1 each day

date = Full date variable

PM25 = Fine particulate matter pollution reading

station_id = A unique ID number for each pollution station

Delhi = A dummy / indicator variable that reads 0 when the station is outside Delhi

1. On a single plot show two lines - one for average pollution over time inside Delhi and the other for average pollution over time outside Delhi. Start the series at 15 December and end on 15 January. Draw a vertical line at Jan 1, the date when the scheme came into force. Do you see any difference in pollution before and after the policy starts? Inside Delhi? Outside Delhi? (**2 point**)
2. Produce a scatter-plot of the *difference* in average daily PM25 for Delhi and outside Delhi. Note: The daily average in Delhi means the average reported by all Delhi stations on that day (**1 points**)
3. Raw data is often messy because variation is generated by many unobserved factors. This is one of the reasons to use simpler models to describe the essential features of noisy data, for examine linear regression models. Run the following regression *using only data from Delhi* and provide the output complete with parameter estimates and standard errors (**2 points**).

$$PM25_{st} = \alpha + \gamma_s + \beta(Policy_{st}) + \epsilon_{it}$$

In this model α is the constant term, γ_s is a fixed effect for each station in Delhi, s stands for station s, t stands for time, *Policy* is a dummy that is 1 between January 1 and January 15 (both days included) and zero everywhere else. To run this model you will need to create the *Policy* variable yourself.

Explain what this regression model is doing in your own words. What is the meaning of β. Can it be used to reliably evaluate the performance of the policy? Why or why not and under what conditions?

4. Now run the following regression using only data from Delhi (**2 points**)

$$PM25_{st} = \alpha + \gamma_s + \beta(Policy_{st} \times time) + \epsilon_{it}$$

Here *time* is the continuous variable in the dataset counting days and everything else is as above. The expression $a \times b$ in this equation stands for three terms $a + b + a \cdot b$ where $a \cdot b$ is the product of a and b. You can create the product variable manually and write out these terms yourself, or you can let R do it for you. Putting *Policy*time* into your regression formula will automatically tell R that you want all three terms. You should end up with the following parameter estimates: Constant term, fixed effects for stations, 3 'betas' corresponding to $Policy$, $Policy \cdot time$, and $time$.

Explain what this regression model is doing in your own words. What is the meaning of the 3 βs. Can the model be used to reliably evaluate the performance of the policy? Why or why not and under what conditions?

5. Another way of trying to evaluate the effect of the driving rationing policy is to compare places with and without the policy. Use 'part2.Rdata' to estimate the following model. (**2 points**):

$$PM25_{st} = \alpha + \gamma_t + \beta(Delhi \times Policy_t) + \epsilon_{it}$$

Here γ_t is a fixed effect for every day and *Delhi* is the dummy that is 1 for stations inside Delhi and 0 outside. As before, $a \times b$ in these regression equations stands for three terms $a + b + a \cdot b$ where $a \cdot b$ is the product of a and b. Putting *Policy*Delhi* into your regression equation will automatically tell R that you want all three terms.

Explain what this model is doing and what the different parameters mean (**2 points**).

6. Estimate the difference-in-differences model below. *Here we will again use 'part1.Rdata', which has hourly readings from each station.* As before, explain what each of these coefficients is doing and whether you can use any of them to evaluate the performance of the policy. Are there any conditions under which this would not be a good way of proceeding? (**3 points**).

$$PM25_{sht} = \alpha + \gamma_t + \omega_s + \theta_h + \beta(Policy_{sht} \times Delhi) + \epsilon_{sht}$$

In the equation:

- $PM25_{sht}$ is the particulate matter reading from station s on hour h, on day t. h ranges from 1-24, so in general for every station there are 24 observations each day (if there is missing data there may be fewer than 24 observations).
- γ_t is a fixed effect for every day. This controls for any factors that effect pollution on any given day for all monitors.
- ω_s is a fixed effect for every station. This controls for any systematic differences across monitors that do not change with time.
- θ_h is a fixed effect for every hour. This controls for systematic hourly differences that effect pollution everywhere. For example pollution may always be high at 4pm.
- $Policy_{sht}$ is similar to the variable we used previously except we will make it more precise. Earlier we used daily data and defined $Policy_{st} = 1$ on days when the odd-even was in force. But in fact, the policy was only enforced between 8am and 8pm. So $Policy_{sht}$ should be 1 only between 1 January and 15 January, *and* only between 8am to 8pm (both inclusive). Everywhere else $Policy_{sht} = 0$. You will need to create this yourself.

Note: You may find that when you run this model R reports coefficients corresponding to $Policy_{sht}$ and the product $Policy_{sht} \cdot Delhi$ but leaves out the third term, $Delhi$. If that happens, don't worry about it, we are interested only in the first two. Maybe you can guess why it might happen (no points hanging on this answer).