Problem Set 1 - PANEL

Professor: Chiara Monfardini;
Teaching assistant: Vito Stefano Bramante

- This problem set was sent on **February 25, 2022**
- Due date: **March 6, 2022** at **midnight**
- E-mail your assignment at vitostefano.bramant2@unibo.it no later than **midnight**. Assignments sent after the deadline will not be graded (no exceptions).
- You must attach a single zip file containing: (i) a pdf answer sheet; (ii) the Stata log-file; (iii) the Stata do-file. The pdf file should be **no longer than 5 pages (w/out tables).**
- Each table and graph in the pdf file should be fully reproducible. **By simply running the .do file** I should be able to reproduce the **exact** table or graph you are showing in the pdf, including title, variable names and numbers. In the .do file you have to signal clearly which chunk of code reproduces which table or which chart. **Doing so will guarantee you up to 2 bonus points on each problem set**.
- Name the zip file (and each file) as *surname*1_*surname*2_*surname*3.*zip*; remember to write the name, surname and student number of each student in the answer sheet.
- The grade for the 3 assignments will be 40% of the final grade.
- Please follow carefully the instructions detailed above. **Any misconduct will negatively impact the grading of the assignment.**

Social Capital is defined as the variety of social interactions, networks, and groups that link people in society together. In his 2009 paper Benjamin Olken investigates whether the expansion of television and radio programming can threaten social capital formation, a hypothesis put forward by the sociologist Robert Putnam a few years earlier. To do so he explores a country, Indonesia, and a time-frame, 1991 to 2003, in which the supply of TV channel increased from one public channel to eleven public or private channel. As a measure of social capital, Olken uses participation to social organization. He documents how each additional television channel introduced corresponds, on average, to a decrease in citizens' participation in social organization. In this exercise you are asked to replicate these findings [1].

Crucially for his purpose, prior to 1993, the only existing TV channel was a public TV station that could broadcast to all locations in Indonesia. In 2003, instead, while the total number of TV stations that could broadcast increased to eleven, not all villages could receive their signal equally. Different villages might receive a different number of TV stations, creating, from a researcher's standpoint, enough variation to assess how, starting from 1990, social participation has evolved differently in locations with different availability of TV channels.

According to Putnam's hypothesis, locations receiving a higher number of TV channel should, on average, experience a reduction in social capital. In practice, this translates to regressing social participation onto the number of TV channels available and to expect the coefficient to be negative and statistically significant.

- In the first part you are asked to explore this hypothesis in a more descriptive fashion. Is there a reduction in social participation in the time period 1990-2003? Is there an increase in TV channels? How widespread is the availability of TV channels across different geographical location?
- In the second part you are asked to think carefully about the framework useful to address the hypothesis under investigation. What is the unit of analysis we should use? What's the contribution of the panel dimension in the data?
- In the third part you are asked to perform the analyses and interpret the results.

[1] While investigating television and radio seems history in 2022, new forms of media are replacing them with an offer that has unprecedented variety and scope across the globe. If you are interested in this contemporary perspective a good starting point is this blog post

Part 1: Data Exploration

In this first part we will take a closer look to our variables of interest. This part also allows us to play around with the data and gain more confidence. Let's start with our dependent variables.

1. Create a new variable called *"org"* with value 1 if each individual participates to at least one social organization, 0 otherwise. Be careful when handling missing values. The resulting "org" variable should have the same umber of missing value than *any* of the other social participation variables.

2. For each type of social organization (including the one you just created in point 1) plot the average participation in 1991 and in 2003. This will be your **Graph 1** and it can just be a simple bar plot. Briefly describe your findings. Do you observe a decrease for *every* organization? Is what you observe consistent with Putnam's hypothesis? Do you think the analysis is worth pursuing (be honest with your answer, you'll have to go further anyway!)?

3. Variabilities are measures of spread in a distribution. Between and within-variability are no exceptions. Using the following formula compute, and report, the between variability for each of the five variables of social participation:

$$stdev(\bar{x}_{sd})$$

where $\bar{x}_{sd}$ is the subdistrict-level mean of the variable x under consideration. Plot the $\bar{x}_{sd}$ to take a look at the distribution [2] . This will be your **Graph 2**. Ideally, you should combine the graphs of all variables into 1, using the command `graph combine`.

4. To compute the within-variability, instead, use the following formula:

$$stdev(x_{i,sd,t} - \bar{x}_{sd} - \bar{\bar{x}})$$

where $x_{i,sd,t}$ is a single observation for each individual, $\bar{x}_{sd}$ is the subdistrict-level mean of the variable x under consideration and $\bar{\bar{x}}$ is the overall mean of the variable x under consideration. Report them.

5. Let's move to our main covariate, the number of TV channel. Compute the beteween and within-variability, and reproduce the same plots of point 3 (on the between variability). This will be your **Graph 4**. Do you think is it worth proceding with the analysis now? Why?

Part 2: Conceptual framework

Let's go back to our original dataset[3]. Consider the baseline POLS specification (forget about any type of fixed-effect or dummy) where you only regress the participation to a social organization onto the number of TV channels. Write down the equation, estimate it and interpret the coefficient of the number of TV channels. Store the results in **Table 1**. Remember that the data we are working with are at the individual level. Is the coefficient consistent with the theory?

1. Consider now our dataset. Despite being an individual-level dataset, our main regressor, the number of TV channels is at the subdistrict level, meaning that individuals living in the same subdistrict might participate differently to social organization but all share the same number of TV channels. To get a better grasp, assume individuals that live only in two different subdistricts. Let's call them "*subdistrict A*" and "*subdistrict B*". Write down the specification including subdistrict dummies (**you don't need to estimate it!**). Do you need a dummy for both subdistricts? What is the role of the constant? Interpret each coefficient *[Hint: You should have two, in total].*

2. Consider now to have data for two different waves at two different time periods $t = 0, 1$. Augment the previous specification introducing dummies for the two waves (**you don't need to estimate it!**). Do you need them both? What is the role of the constant now? Interpret each coefficient *[Hint: You should have three, in total]*..

3. A lot of things can change in a decade. How can we be sure that is actually TV to have a negative impact on social participation? Ideally, we would want to control, for each subdistrict, for other unobservable variables that, like the number of TV channels, change over time. In order to do so we would, ideally, interact the subdistrict dummies with the wave dummies. Try it: augment the previous 2-subdistricts/2-waves specification interacting the subdistrict dummies with the wave-dummies. What happens to the number of TV channels? Can we estimate it? (**you don't need to estimate it!**) *[Hint: Is the number of TV channels at the individual level or at the subdistrict level?].* One possible alternative way to control for location-specific trends is given in the next part.

[2] One way to do so is by using the command `collapse` to create a new dataset containing only the $\bar{x}_{sd}$. After that you can create a new variable that ranks the $\bar{x}_{sd}$ in decreasing order and run a twoway plot with the $\bar{x}_{sd}$ and the (reverse) rank variable on the x-axis.

[3] This means that all the analysis in part 1 should be within a *preserve-restore* sequence or should be saved for further analysis

Part 3: Analysis

Consider as your main specification:

$$SP_{sd,i,t} = \alpha_{sd} + \alpha_t + \delta TV_{sd,t} + \beta X_{sd,i,t} + \epsilon_{sd,i,t}$$

where $SP_{sd,i,t}$ is participation to a social organization, α_{sd} are subdistrict (Kecataman) fixed-effects, α_t are wave dummies, $X_{sd,i,t}$ are age, gender, years of education and log expenditure per-capita.

1. Focus only on the vairable *org*. Estimate the specification using the command *reg*. Cluster at the subdistrict level. Interpret the coefficients and their statistical significance. What is the role of the subdistrict dummies?

   Using the command `collapse`, create a new dataset averaging by subdistrict (kecnum) and wave.

   Use the command `xtsum` to explore the within and between variability. Run the same specification as in question 1 but this time use the command `xtreg`. In one table show the results obtained using the command `reg` and the ones obtained using the command `xtreg`. This will be your **Table 2**. You can omit the dummies coefficients but you have to include the coefficients and the std errors of your main regressors. What do you find? Are the coefficients coming from the two regression the same? Do they tell a consistent story? Is the story consistent with the theory?

2. In part 2 question3 we have discussed the impossibility to control for unobserved variables that, like the number of TV channels, might change for each subdistrict. One way out is to expand the goegraphical dimension and interact it with time. Run the specification below using both `reg` and `xtreg` and compare the results. This will be your **Table 3**. One again, you can omit the dummies coefficients but you have to include the coefficients and the std errors of your main regressors. Compare the results. Do they tell a consistent story? Is that story consistent with the theory?

   $$SP_{sd,i,t} = \alpha_{sd} + \alpha_{d,t} + \delta TV_{sd,t} + \beta X_{sd,i,t} + \epsilon_{sd,i,t}$$

   where $\alpha_{d,t}$ are district*wave dummies. By introducing them we are creating district-specific trends, that while not accounting for time-varying unobservable characteristics for each subdistrict, allow us to do so for each district.

Codebook

The dataset contains a mixture of individual-level survey data and subdistrict-level data. The variables are:

- **soyouth**: coded 1 if individual participated in any youth organization in the last three months
- **soreligious**: coded 1 if individual participated in any religious organization in the last three months
- **sowomen**: coded 1 if individual participated in any women's organization in the last three months
- **soburial**: coded 1 if individual participated in any burial organization in the last three months
- **tvchannels**: Number of TV channels received in kecataman (subdistrict)
- **kecnum**: Kecataman (subdistrict) code
- **wave**: survey wave (1991 or 2003)
- **kabidwave**: code of the district (kabid) wave combination
- **age**: Age of the individual
- **gender**: Gender of the individual
- **years_educ**: Years of education of the individual
- **lnexpcap**: log expenditure per-capita