Problem Set 4
Quantitative and Statistical Methods II

General Instructions

- The problem set is due by March 7th at 23h59;
- You should send it by email *bruno.conte@barcelonagse.eu*; it must include your *unique* Stata code, *a unique* log file, and your answer sheet. Tidiness is appreciated – e.g. material correcly labeled and organized in a zip file;
- Working in teams is allowed and strongly recommended (keeping the same groups as in the presentations is encouraged);
- The datasets needed to solve the computational questions are uploaded in the Classroom material.

Part 1: "Minimum Wages and Employment: a Case Study of the Fast Food Industry in the New Jersey and Pennsylvania", by David Card and Alan Krueger, 1993

In this paper, the authors evaluate the impact of a particular policy - increase in minimum wages - on Full-Time Equivalent Employment rates using a Differences-in-Differences strategy. Read the introduction and empirical strategy sessions to understand well the differences between the New Jersey and Pennsylvania states, and what the policy was about in details. Look at the `class4.do` file covered in class for more details on its practical estimation. The data you will need for the following exercises are on `cardkruegerdata.dta`.

1. State which are the necessary assumptions for the DiD strategy. Interpret them in the context of the paper. Do you think there is any concern regarding its validity in this context?

2. Compute the average Full-Time Equivalent employment in NJ and PA stores before and after the policy change (occurred in April 1992). Use these estimates to compute the Differences in Differences estimator

$$\hat{\beta}_{DD} = (\mathbb{E}[Y_{it}|D_i = 1, t = 1] - \mathbb{E}[Y_{it}|D_i = 1, t = 0]) - (\mathbb{E}[Y_{it}|D_i = 0, t = 1] - \mathbb{E}[Y_{it}|D_i = 0, t = 0]).$$

   Comment the results. **[Hint: Use the Stata command *collapse*, as done in class. D_i, here, is a dummy for the treated state, and *t* for the time period after the policy was implemented.]**

3. Write the corresponding regression model that will lead the same Differences in Differences estimator obtained above. Implement it in Stata. Make sure results are identical to what you have just obtained.

Part 2: "Dynamic Inefficiencies in an Employment-Based Health Insurance System: Theory and Evidence", by Hanming Fang and Alessandro Gavazza, American Economic Review, 2011

In this paper, the authors empirically verify their model of misallocation of investment in health care due to frictions in the labour market. They use a panel data on health care expenditures together with longitudinal information on individuals' job tenure to assess the effects of job turnover on lifetime investment in health and medical expenditures when retired. Read the text carefully to understand in details its context and the author's empirical strategy to answer the questions below.

4. Define y_{it} as the medical outcome of individual i at year t. Then consider the following static model

$$y_{it} = \beta_0 + \beta_T \log(\text{Job Tenure}_{it}) + \underbrace{\eta_i + v_{it}}_{u_{it}}, \tag{1}$$

   where η_i capture individual fixed characteristics that are unobserved. If you were to estimate this model with a classic OLS regression, which would be your concern regarding the relationship between Job Tenure$_{it}$ and η_i? In other words, which is(are) the assumption(s) you would need to make sure $\hat{\beta}_T$ is consistent? Are you on a Fixed or Random Effects scenario?

5. Let y_{it} be "Doctor Visits". Using `paneldatafangcavazza.dta`, estimate the trivial OLS model above, but including a dummy for each $year - region$ and the covariates the authors control for. You will be replicating the first column of Table 3 in the paper, so check which are those covariates there. **[Hint: make sure things work: the covariates you use here are going to be needed in many other estimations that follow!]**

6. In a Random Effects scenario, in which $\mathbb{E}[\text{Job Tenure}_{it}\eta_i] = 0$, a simple OLS regression will be subject to an efficiency issue. The reason is that the unobserved error component, $u_{it} = \eta_i + v_{it}$, suffers from serial autocorrelation, i.e. $\mathbb{E}[u_{it}u_{is}] \neq 0, \forall t \neq s$.

   a. One way for such correlation is through the *Feasible GLS* estimator. Describe how this estimation is done and the necessary assumptions.

   b. Using `paneldatafangcavazza.dta`, estimate such model. You will be replicating column 2 of Table 1 below. Make sure you use the same controls/dummies you used before. **[Hint: recall to use the *xtreg ..., re r* command. Do not forget to set your data as panel, i.e. *xtset ...* as done in class.]**

7. Now suppose we assume that $\mathbb{E}[\text{Job Tenure}_{it}\eta_i] \neq 0$, i.e. a Fixed Effects scenario.

   a. Explain in words why, in the context of the paper, that must be the case. Is that consistent with what you answered in question 1?

   b. One way of estimating a Fixed Effects model is the *First Difference LS* model. Describe how this estimation is done, the necessary assumptions, and estimate it with the dataset used so far. Make sure the results match with column 3 of Table 1 below.

   c. Analytically show why this estimator solves the unobervables' problem of the estimates in question 5 and 6. That is, show that $\mathbb{E}[\Delta x_{it}\Delta u_{it}] = 0$ ($x_{it} \equiv \text{Job Tenure}_{it}$).

   d. Another way doing inference in the current context is through the *Within Group* Estimator. Describe which are **the two ways** of doing this estimation. Explain how they are done and the necessary assumptions. How do these assumptions differ from the ones from the *First Difference LS* model?

   e. You will now use the `paneldatafangcavazza.dta` data for doing **both types** of the WG estimation. For doing so, **aggregate the data at the region-year level**, so now the units of observation are averages at a region, in a certain year. Use the *collapse* command as shown in class. Disregard the region-year dummy (`y_r`) variable used before. Make sure the results are identical between them and that they do match column 4 of Table 1 below.

8. Now suppose that doctor visits have a persistent behaviour, i.e. individuals that go often to the doctor would go more in future periods. Let us forget for a moment about the Job Tenure$_{it}$ regressor. Then, we can re-write model (1) as

$$y_{it} = \alpha y_{it-1} + \underbrace{\eta_i + v_{it}}_{u_{it}}, \quad |\alpha| < 1 \tag{2}$$

   α is the *persistence* parameter, which tells us how past values of the dependent variable is associated with its present values. This is a *Dynamic Panel* framework.

   a. State which are the necessary assumptions for estimating the model (2) above.

   b. Why should $|\alpha| < 1$?

   c. We saw in class that, by construction, $\mathbb{E}[y_{it-1}\eta_i] \neq 0$. Does that make sense in the context of the paper? Comment it in a few words.

   d. We know that OLS and RE would be biased in this context. Suppose one tells you to estimate it with the *First Difference LS* model. Show that, in this case, estimates are also biased. **[Hint: you are asked to show that $\mathbb{E}[\Delta y_{it-1}\Delta u_{it}] \neq 0$.]**

   e. We also saw in class that Anderson and Hsiao propose an IV approach to estimate a model like (2) in first differences. Arellano and Bond, instead, provide an alternative GMM approach (aka *First Difference GMM*). Describe this method in details. **[Hint: answer it in words, without heavy algebra. The most important here is to make sure you understand it!]**

f. Arellano and Bover (1995) improved the GMM method above to what is known as *System GMM*. Describe this method, emphasising the different assumptions from above. **[Hint: the same hint.]**

g. Use the data `gmmdatafangcavazza.dta` to replicate column 5 of Table 1 below (or 4 of Table 3 in the paper, aprox.). You must use the same specification (dependent variable, regressors) used in exercise 2, but adding the lag of the dependent variable as a regressor and the level and lags of the exogenous regressors. **[Hint: look carefully at *xtabond2*'s help in Stata to understand in more details how System GMM is implemented in Stata. Moreover, check the `class4.do` file used in class for other hints!]**

Table 1: Results from Estimations of models (1) and (2).

Estimate	Pooled OLS	Random Eff.	First Diff.	WG (region-year level)	System GMM
$\log(\text{Job Tenure}_{it})$	-.0034***	-.0036***	-.0055***	.0605	-.1009*
	(.0007)	(.0008)	(.0015)	(.0347)	(.0514)