Microeconometrics
Spring Semester 2019/20
Problem set 3
Due: May 5^{th}

Note: Send a copy of your solutions to Sónia Félix at sfelix@novasbe.pt and to me (teresa.molina@novasbe.pt). The solutions should include a .pdf or .doc document with the answers to each one of the questions ***and*** *the log files or do files of the exercises done using stata. Late problem sets will not be accepted.*

Exercise 1 (Panel data models)

For this exercise please create a do-file and corresponding log-file to hand in with your solution.

Consider a subset of the data used by Vella and Verbeek (1998) *wagepan.dta* to estimate the effects of unions on workers' wages. The dataset is comprised of 545 men who worked in every year from 1980 through 1987 in the United States.

Consider the following model:

$$lwage_{it} = a_i + \theta_t D_t + \beta x_{it} + \pi z_i + u_{it}$$

where i denotes the worker and t the year. The vector x_{it} is comprised of *exper* (labor market experience) and its square *expersq*, *married* equals 1 if the individual is married, and *union* equals 1 if the worker is unionized. The vector z_i includes the variables *black* equals 1 for blacks, *hisp* equals 1 for hispanic workers, and *educ* denotes the number of years of education.

(a) Explain which effects parameters θ_t and a_i are likely to capture.

(b) If unions are successful in their wage negotiations with employers, what should be the sign of β_{union}?

(c) Estimate the equation by pooled OLS. Do you find any evidence for a union effect? Are the assumptions required for these estimates to be consistent plausible? If not, what would be the asymptotic bias you would expect in the union estimate?

(d) Now estimate the model in first differences (FD). Can we estimate the returns to education in FD? Why? What about race effects and experience?

(e) Comment what the FD results of the union effect suggest on the correlation between union and a_i.

(f) Considering the time-varying variables, estimate the equation using the within estimator and the random effects estimator. What are the necessary assumptions for consistency of the random effects estimator?

(g) Perform an Hausman test for fixed effects. What do you conclude? Does the random effects estimator or the pooled OLS provide consistent estimates? State clearly the null and alternative hypotheses, the test statistic, and the critical region.

Exercise 2 (Differences in differences)

The goal of this exercise is to apply some basic difference-in-difference estimations.

The specific intervention we will be analyzing is the construction of new secondary schools in Wonderland. In Wonderland, there were some communities that had new secondary schools, and other communities that did not- this leads to the variation necessary to apply difference-in-difference. In short, we want to see if individuals who lived in areas with new secondary schools completed more school than individuals who lived in areas without secondary schools.

We will be looking at individuals who live in communities, so there are some individual level variables refer to individual characteristics, such as gender, age, education and some community level variables refer to characteristics of the community, such as access to clean water, electricity, and paved roads. Communities where a new secondary school was built will be known as "treatment" communities. Communities where no secondary school was built will be known as "control" communities. We will also look at two cohort groups. Young cohorts (aged 6-16 in 1985) and old cohorts (aged 21- 41) in both treatment and control communities. The idea is that new secondary schools should only affect young people who are still in school. If you have completed your studies, a new secondary school in your community will not change how much education you get. The idea is that the "treatment" or new secondary schools should only affect the young cohort living in treatment communities. This generates the difference- in-difference design.

1. Present individual summary statistics for the study sample for treatment communities vs. control communities. Do a t-test to see whether differences in age, education, and gender are statistically different between the treatment and control group. Do you see any differences? Are you concerned by any of the differences? How could they affect the analysis? (hint use: `ttest (variable), by(treat)`
2. Do a t-test to see whether differences in access to electricity (electric), piped water (pipwater), and distance from the capital (distcapital) are statistically different between the treatment and control group at the community level . Do you see any differences?
3. Based on your answer to either question ~~4 or 5~~, do you think that treatment communities are different from control communities? 1 and 2
4. Run the following regression $Primary_i = \alpha + \beta T_i + u_i$, only for the young cohort (ycohort==1). Based on your estimates, can we say that constructing new secondary schools has a direct and causal impact on primary school completion? Why?
5. Do the regression again but add additional controls. Explain any difference with previous results.
6. We are now just going to get the mean value for each cohort group and complete the following table: Replicate table using sample means

	Treatment communities	Control communities	1st Diff
Young Cohort	.9016393	.7942029	0.1074364
Old Cohort	.4285714	.3764259	0.0521455
1st Difference	0.4730679	0.417777	0.0552909

7. Now, estimate the treatment effect using a standard diff-in-diff regression

$$Primary_i = \alpha + \beta_1 T_i + \beta_2 YC_i + \beta_3 (T^* YC) + u_i$$

How does it compare your estimate to the treatment effect obtained in the table?

8. Based on your analysis do you think building new secondary schools is effective at increasing primary school completion rates? What are some potential problems with the above analysis?

Exercise 3 (Reviewing LATE)

Say children can enroll in child care or not, and denote enrollment as $D_i = 1$ and non-enrollment as $D_i = 0$. Several researchers and politicians argue that child care may have positive effects on childrens cognitive development, while others argue that it has negative effects, particularly on young children. Let y_i be child is performance on a language test at age 7.

1. The difference-in-means estimator $\bar{d} = \bar{y_1} - \bar{y_0}$, compares the observed means in the subpopulations with $D_i = 1$, $\bar{y_1}$, and with $D_i = 0$, $\bar{y_0}$. Show that $\bar{d}$ is a biased estimator for the average treatment effect on the treated, and derive an expression for the bias. What sign do you think the bias takes in this application? Explain. (*Hint: Take expectations* $\rightarrow E(y_1 - y_0)$)
2. Say you get access to a program that randomly allocates a subset of child care places to children on the waiting list. Let $Z_i = 1$ if the child receives an offer, and $Z_i = 0$ if the child does not receive an offer.
   (a) Who are the compliers in this application? Who are the always-takers and never-takers? Do you think there might be defiers?
   (b) Under what conditions is the local average treatment effect (LATE) for the complier group identified?