Problem Set 2

EE468 Integrated Public Economics, Development and Political Economics

Due in class on April 4. All questions count equally toward 10% of your grade. You are encouraged to work together. But make sure to submit your individual work.

Instruction: You need to download Stata datasets attached in Wattle. You can hand in the log file especially the parts that are directly relevant for answering questions.

1. Worms in Kenya

The dataset, ted_miguel_worms.dta is from Ted Miguel's (UC Berkeley) and Michael Kremer's (Harvard) de-worming project in Kenya. Make sure to read the paper to familiarize yourself with the experiment before answering this question.

Miguel and Kremer randomise over schools (not individuals) and to introduce de-worming drugs to randomly selected treatment group to estimate the effect of deworming on school attendance.

1.1 First, why randomize at the school level? Explain what issues might arise in your evaluation if you randomize at the individual level.

1.2 Suppose you had pre- and post- treatment school attendance records for all schools, describe the calculations/comparisons you would do in order to estimate the effect of deworming on school attendance. Hint: think of difference in difference (DD) method and describe this in the form of DD grid or diagram. Why might you prefer this DD estimate to a straightforward OLS estimate using only post-treatment data (i.e. mean attendance in program schools minus mean attendance in non-program schools after the program)?

1.3 Write down the regression you want to run in to estimate the DD you described in part (b)? Explain the meaning of each coefficient.

Now look at the data set: ted_miguel_worms. Use the *describe* command to see the meaning of available variables. The authors have pre- treatment health, but only for the Group 1 kids - since it was considered unethical to get health information for control group kids -and thus this data cannot be used in regressions. So we cannot run the DD proposed above

1.4 Describe the data: How many observations are there per pupil? What percentage of the pupils are boys? What percentage of pupils took the deworming pill in 98 and 99? What percentage of schools was assigned deworming in 98? Is this more or less than the percentage of pupils who actually took the pill in 98? Why might this be so?

1.5 Using the data: summarise infection outcome (post 1998 treatment) of pupils in group 1 (in 1998 treated school) and those in group 2 (1998 control school). Are the outcomes from two groups different and are these difference statistically significance? (Hint use *ttest* or *ttesti* command). Explain what might this difference represent? Is this a good estimate of the health effect of de-worming program? Think about what effect might this not include.

1.6 Explain in word, how the authors proceed to resolve the above issue and so to estimate the overall health effect of de-worming program (Hint: explain how they incorporate cross-school externality into the overall estimate).

2. Estimating impacts of INPRES school construction program and return to schooling in Indonesia

The dataset, supas.dta, is from Ester Duflo's 2001 AEW paper on schooling and labor market outcomes of school construction program (INPRES) implemented in 1973-74 and 1978-79 in Indonesia. The data contain those born between 1950-1972 (age 2-24 in 1974 with those age 2-12 in 1974 being young enough to be exposed to INPRES)

Because this data set is 2.5Mb while the Stata default memory setting is 1Mb, begin your Stata session by typing "*set mem 20m*"; you will also need to type the command "*set matsize 800*"). Use the *describe* command to see the meaning of available variables. Spend some time understanding the notation and data well, and the methods are explained step by step.

2.1 Describe the data: summarise years of schooling and lnwage of the whole sample. What percentages of individuals were born between 1968-1972 (age 2-6 in 1974, call this "young"), born between 1957-1962 (age 12-17 in 1974, call this "old"), born between 1950-1956 (age 18-24 in 1974, call this "veryold")? What percentages of individuals are in the high program intensity region? Is there difference of number of schools for 1000 children in the high Vs. low program intensity region? If so, is the difference statistically significant?

2.2 First, estimate the impact of INPRES on schooling using simple difference-in-differences (DD) with cross-cohort (young versus old) and cross-regional (high-intensity versus low-intensity) variation.

(i) What are the mean years of education of the young cohort in the high- and low-program intensity regions; are these significantly different? What are the mean years of education of the old cohort in the high- and low-program intensity regions; are these significantly different? Interpret these results.

(ii) Are there significant differences between mean schooling of the young and old cohorts in each of the program intensity region? Are these increases in education between cohorts at the same rate in both program regions? What might be accounted for these differences across regions? (Hint: as the young cohort was exposed to the program, the larger increase in education between cohorts in the high-program intensity region should be a result of the program)

(iii) Bring these results together to construct a DD two-by-two grid (similar to panel A in Table 3 of the paper). What might be the effect of INPRES on education according to DD? Interpret the result, how large would years of schooling increase as a result of the program. You also learn from (ii) that the difference between the number of schools constructed per 1,000 children in high and low program regions is 1.04). What does this imply about an increase in years of schooling as a result of one additional school constructed for 1,000 children?

(iv) Under what assumption would this estimate be an unbiased estimate of the INPRES impact? Redo the above estimates and construct a DD grid for the "control experiment" among old and veryold cohorts (who both were not exposed to the program). What do you find? Can these results confirm the identification assumption above? (Hint: as both cohorts were not exposed to the program, no significant difference in the increase in education between cohorts across the two regions implies that the only systemic factor driving these cross-region differences is the program.)

2.3 We will now try to estimate DD effect of INPRES on years of schooling using a simple linear regression:

$$yeduc_{ijt} = c + \alpha_j + \beta_t + (P_j \cdot T_i)\gamma + (C_j \cdot T_i)\delta + \varepsilon_{ijt}$$

where α_j is region of birth fixed effect, β_t is cohort of birth fixed effect, P_j reflects intensity of exposure of the program region, T_i is dummy indicating whether individual belongs to exposed cohort (young) and C_j represents other control regional-specific variables.

(i) Explain this regression function, what could these coefficients represent?

(ii) First, we will replicate result in (2.2) by estimating the above equation in a subsample restricted of only those born in young or old cohorts. We will use simple cross-cohort (young versus old) and cross-regional (high-intensity versus low-intensity) variation and control only for simple region (high, low) and cohort (young, old) fixed effects. You can do so by the following command

- *gen young_high = young*high*
- *regress yeduc high young young_high if young==1|old==1*

Interpret the estimate. The two fixed effects are the same as simple region/cohort dummies in this simple case. You should get the same estimate as (2.2).

(iii) Let's explore and estimate various versions of this equation. Run each of the following command, explain your results. Under what assumption is this an unbiased estimate of the effect of INPRES on education?

Use full region of birth fixed effect α_j instead of simple (high, low) dummy

- *xtreg yeduc young young_high if young==1|old==1, i(ROB) fe*

Use number of schools for 1000 children to represent program intensity P_j instead of just a simple (high, low) dummy and adding a region control of number of children in 1971 as C_j

- *gen young_intensity = young*prog_int*
- *gen young_ch71 = young*ch71*

- *xtreg yeduc young_intensity young young_ch71 if young==1|old==1, i(ROB) fe*

2.4 We will now estimate the effect of INPRES on years of education using a more generalised linear regression:

$$yeduc_{ijt} = c + \alpha_j + \beta_t + \sum_{l=2}^{12} (P_j \cdot d_{il})\gamma_l + \sum_{l=2}^{12} (C_j \cdot d_{il})\delta_l + e_{ijt}$$

This allows us to extend time dimension of the individual exposure to the program (written as a simple dummy T_i whether one is in young cohort in the earlier equation) to include age-specific cohort. Those who were age 2 to 12 in 1974 are young enough to have been exposed to the program. d_{il} thus represents dummy variables whether individual i is age l in 1974, where l = 2,…, 12. This generalized form thus allows us to estimate program impact on a given age cohort (γ_l with l from 2-12) using age 13-24 as our control group.

To estimate this, we need to first generate a dummy variable for those 2-12 years in 1974 (born between 1962-1972) equal to 1 if the individual was born in that year and 0 otherwise by using command *gen d62 = (YOB==62)* and redo this command for *d63, d64,…* up to *d72*. We then need to interact each and every dummy with program intensity (by gen *d62_intensity = d62*prog_int* and redo for all *d63,…, d74*) and region control (by *gen d62_ch71 = d62*ch71* and redo for all *d63,…, d74*)

You can launch these commands to generate these variables all in once by cutting and pasting the following loop command to your command window

- *forvalues i = 62(1)72 {*
  gen d`i' =(YOB==`i')
  *gen d`i'_intensity = d`i'*prog_int*
  *gen d`i'_ch71 = d`i'*ch71*
  }

Now, we are ready to re-estimate the impact of INPRES on years of schooling by age cohorts by the following command

- *xtreg yeduc d62_intensity d63_intensity d64_intensity d65_intensity d66_intensity d67_intensity d68_intensity d69_intensity d70_intensity d71_intensity d72_intensity d62 d63 d64 d65 d66 d67 d68 d69 d70 d71 d72 d62_ch71 d63_ch71 d64_ch71 d65_ch71 d66_ch71 d67_ch71 d68_ch71 d69_ch71 d70_ch71 d71_ch71 d72_ch71, i(ROB) fe*

We can also test the hypothesis that coefficients of the exposure variables (interactions between age cohort dummies and program intensity in birth region) are jointly zero by launching the following command right after above:

- *test d62_intensity d63_intensity d64_intensity d65_intensity d66_intensity d67_intensity d68_intensity d69_intensity d70_intensity d71_intensity d72_intensity*

Interpret the results. Variation of program impacts by age cohorts should follow those in Figure 1 of the paper. Duflo documented that on average, 1.98 schools were built per 1,000 children. How would you interpret total effect of INPRES on increase in years of schooling for the 2-year-old cohort in 1974 with this information?

2.5 Use the same generalized regression to estimate the effect of INPRES on log hourly wage (by replacing *yeduc* with *lhwage* in the above Stata command).

Interpret the results. Plot impacts of INPRES on log wage and years of schooling by age cohorts. You should get similar pattern as Figure 3 in the paper: the program effects on wage track well those on schooling. As Duflo interprets, this result implies that the program effect on wages was caused by the changes in years of education. She then proceeds to use individual exposure to INPRES program as exogenous instrument to estimate return to education in Indonesia.

2.6 Let's now shift to estimate returns to schooling. Recall the wage-setting equation from the education model discussed in class: $lnw_i = a_i + bS_i$, where w_i is earning and S_i represents years of schooling.

(i) Interpret the parameter b in this equation. Why might the coefficient on an OLS regression of log wages on years of schooling not be a good estimate of b?

(ii) We will now employ instrumental variable technique to try to get an unbiased estimate of b via 2SLS. We want to estimate the effect of years of schooling on log wages, but years of schooling is an endogenous variable. The idea behind instrumental variables is to use other variables, called instruments, which generate exogenous variation in years of schooling. In our case, we will use "exposure to the INPRES program" captured by $\sum_{l=2}^{12}(P_j \cdot d_{il})$ as instruments that will generate exogenous variation in years of schooling.

The required properties of good instruments are (a) highly correlations with schoolings and (b) no direct effect on log wages. From the past results in 2.4 and 2.5, would these be good instrument, explain? (Hint: look at F-test of overall significant level of instruments in 2.4 and patterns of estimated coefficients of program impacts on wage and education in 2.5)

(iii) To calculate the 2SLS estimate of b:

<u>First stage</u>: run an OLS regression of years of education on the instruments and control variables, and form the predicted values of education from this regression. This is simply what we have done in 2.4. Then,

<u>Second stage</u>: run an OLS of log hourly wages on this predicted value of years of education and the control variables. The coefficient on the predicted years of education is the 2SLS estimate of b.

Compute and interpret the 2SLS estimate of b using the following command. Interpret the results.

- *xtreg yeduc d62_intensity d63_intensity d64_intensity d65_intensity* *d66_intensity d67_intensity d68_intensity d69_intensity d70_intensity d71_intensity d72_intensity d62 d63 d64 d65 d66 d67 d68 d69 d70 d71 d72 d62_ch71 d63_ch71 d64_ch71 d65_ch71 d66_ch71 d67_ch71 d68_ch71 d69_ch71 d70_ch71 d71_ch71 d72_ch71, i(ROB) fe*

- *predict pdt_yeduc*
- *xtreg lhwage pdt_yeduc d62 d63 d64 d65 d66 d67 d68 d69 d70 d71 d72 d62_ch71 d63_ch71 d64_ch71 d65_ch71 d66_ch71 d67_ch71 d68_ch71 d69_ch71 d70_ch71 d71_ch71 d72_ch71, i(ROB) fe*