ECOM90003: Applied Microeconometric Modelling

Assignment 1

This assignment is based on the paper "The local political economy effects of school construction in Indonesia" by Monica Martinez-Bravo published in AEJ: Applied Economics. The paper analyses the effects of a large school construction program in Indonesia on local governance and public good provision.

Your submission should include your written responses to all questions, and then an appendix including all requested tables, and your Stata code. Tables should be numbered sequentially. All tables should have column headings indicating the question they correspond to. Please keep your written answers brief and to the point, referring to the tables by table number and column heading where appropriate.

1. [8 marks] What public goods does Martinez-Bravo use as outcomes to evaluate the effect of mass school constructions? What impact (positive, negative, none) do you think/expect the school construction will have on these outcomes? Justify your response.

2. [3 marks] Explain why comparing the average value of outcomes before and after the first election in areas with a positive amount of INPRES schools may be problematic if the aim is to identify the causal impact of the mass school construction.

3. [3 marks] Explain why comparing the average outcomes for villages with and without INPRES schools may be problematic if the aim is to identify the causal impact of the mass school construction.

4. [3 marks] What is the empirical challenge in identifying the causal impact of mass education interventions on local governance?

5. [6 marks] The author implements a strategy discussed in class using two different specifications (see equations 1 and 2).
   a) Briefly explain the overall *strategy* <u>and</u> the two empirical *specifications* used to implement this strategy.
   b) What is the main parameter of interest in each?
   c) What is the interpretation of these parameters? What different purposes do they serve? Make sure you address both.

6. [2 marks] What is the key assumption that must be met for this strategy to deliver estimates of causal effects of the school construction program?

7. [4 marks] Is the error term in the data likely to be identically and independently distributed (iid)? If not, why not? Is this a problem for inference? How should the standard errors be estimated? Justify your answer (hint: think of the structure of the datasets).

Questions 8 and 9 ask you to look at variation in the data that will be used to identify the impact of the school construction program on local governance.

8. [8 marks] This question asks you to produce summary statistics using the data provided before we estimate the main treatment effects in Table 3 of the paper by Martinez-Bravo.

a) What is the unit of observation in the dataset? How many villages and years are available in it?

b) How many schools were constructed nationwide as part of the INPRES program between 1974 and 1978? What is the average number of schools in the villages studied by Martinez-Bravo? What is the age range of the children that could attend these schools?

c) What proportion of villages had 0, 1 and 2 INPRES schools by the end of the school construction program? (Hint: look at variable *num_PSINPRES1980*).

d) In addition to the variation introduced by the INPRESS school construction program, what other sources of variation help the author discern between public good increases driven by a more educated labor force versus more educated politicians?

e) What is the minimum age requirement for candidates of village head in Indonesia? In what year does the first INPRES cohort meet this requirement?

9. [5 marks] We will now create a table where we will explore the variation in timing of first elections after 1992. Label this table as "Table 1" in your assignment and refer to Table 1 when necessary.

Step 1. Using the data provided to you, create one column with the *proportion* of villages in the data who had their first post 1992 election in each year between 1992 and 2000. These values should match those in Appendix Table 2, column 2. (hint: look at variable ele1v_post92).

Step 2. Create a second column with the *number* of villages who had who had their first post 1992 elections in each year between 1992 and 2000. (hint: remember that each village is repeated 6 times in the dataset, search for commands such as drop and collapse to help you get the right values but remember that you will need the full data for the rest of the questions).

a) Based on the Table 1 you created, is there variation in the timing of the first election across villages? Which years had the highest and lowest proportion of villages with their first post 1992 election? Is there a pattern in the data (increasing, decreasing, neither)?

b) Now refer to Table 2 in the paper by Martinez-Bravo, is this variation in timing significantly correlated with changes in the number of doctors, health centers or health posts in the pre-treatment period? Why would this be a threat to identification?

Questions 10—12 and 15 ask you to estimate various models using the data provided. Please report (all) results for questions 10-12 in a single table labelled Table 2, and use the following format:

	Table 2. The effects of school construction on public good provision								
	Primary health center in the village			Doctors in the village			Acess to safe drinking water		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Post first 1992 election									
Post x num INPRES schools									
Observations									
Mean of dependent variable									
Year fixed effects	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Village fixed effects	No	No	Yes	No	No	Yes	No	No	Yes

Report the standard errors in parentheses under each coefficient. [Hint: See outreg2 or esttab to output Stata results into excel. Make sure to adjust your standard errors following your response to question 4].

Step 1. Create a variable called *post92* which should take a value of 1 if the year of the survey is greater or equal to the year of the first election, and zero otherwise. To check that you're doing this correctly, make sure that the mean of this variable is equal to 0.3852

Step 2. Create a variable called *interaction* where you will interact the *post92* dummy and the demeaned number of INPRES schools per village (this variable is called *num_dev* in the dataset). (Tip: when interacting a dummy and a continuous variable, its useful to subtract the mean off each unit so that we can interpret the interactions as an increase with respect to the mean).

Hint: Stata has a number of commands to help us deal with regressions that include a large number of fixed effects such as "areg" and "reg2hdfe". I suggest using the "areg" command instead of our usual reg command for question 12 in particular.

10. [5 marks] Regress the 3 <u>binary</u> outcomes in panels A, B, and C of Table 3 on the indicator for the first post 1992 election, the number of INPRES schools and the interaction of both variables that you created earlier (without any other controls). Report these coefficients in columns 1, 4 and 7. What is the interpretation (sign, size and significance) of the estimated coefficients?

11. [4 marks] Next, add indicators for year of survey to the above specifications. Report these coefficients in columns 2, 5 and 8. What happens to our estimated coefficients on *interaction* (sign, size and significance)? In light of this, do you think the specifications from Question 10 can be used to obtain the causal impact of the school construction program? Why?

12. [5 marks] Next, add indicators for village to the above specifications. Report these coefficients in columns 3, 6 and 9. What happens to our estimated coefficients for the *interaction* variable (sign, size and significance)?

13. [2 marks] Are these sizable effects? What is the change in the outcomes relative to the sample mean?

14. [3 marks] Martinez-Bravo finds that these effects are heterogenous. When are the estimated interaction effects larger?

15. [6 marks] Re-estimate the above specification but now, instead of including the post dummy, the demeaned number of schools and the *interaction* variable you created earlier, use the double hash (##) function to interact post92 and num_PSINPRES1980

(the non-demeaned variable indicating whether a village has 0, 1 or 2 INPRES schools). Report these coefficients in a separate table labelled Table 3 using the format below (Tip: combine the ## and i. functions to achieve the desired result, e.g., i.post##i.treatment). What do you conclude from these latest results?

Table 3. Intensity of School Construction			
	Primary health center in the village	Doctors in the village	Acess to safe drinking water
	(1)	(2)	(3)
Post first 1992 election			
Post x INPRES schools = 1			
Post x INPRES schools = 2			
Observations			
Mean of dependent variable			
Year fixed effects	Yes	Yes	Yes
Village fixed effects	Yes	Yes	Yes

16. [3 marks] How does the author address the parallel trends assumption? What does she find? (Hint: refer to question 5).