GLOBAL POVERTY & IMPACT EVALUATION (Fall 2009)
Erick Gong & Garret Christensen
Professor Miguel (Faculty Adviser)

Problem Set 4: Difference-in-Difference Estimation (Updated Nov 11, 2009)

The goal of this problem set is to apply some basic difference-in-difference estimations. We will be using data from the World Bank[1] (Beegle et al. 2006). The data comes from a household survey in the Kangera region of Tanzania. We will be using the data to look at the effects of new secondary school construction on educational outcomes.

The specific intervention we will be analyzing is the construction of new secondary schools in Tanzania during the 1980's. In the Kangera region, there were some villages that had new secondary schools, and other villages that did not – this leads to the variation necessary to do a difference-in-difference research design. In short, we want to see if individuals who lived in areas with new 2ndary schools completed more school than individuals who lived in areas without 2ndary schools.

Please submit a do file, log file, and 1 page write up for the problem set. You are welcome to work in small groups, but please submit your own write up. Written responses should be *brief*. Questions with a "*" indicate that they are optional.

Preliminaries

We will be looking at individuals who live in villages, so in some sense, there are two units of analysis. Individual level variables refer to individual characteristics, such as gender, age, education. Village level variables refer to characteristics of the village, such as access to clean water, electricity, and paved roads.

Villages where a new secondary school was built will be known as "treatment" villages. Villages where no secondary school as built will be known as "control" villages.

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 villages. The idea is that new 2ndary schools should only affect young people who are still in school. If you have completed your studies, a new 2ndary school in your village will not change how much education you get. The idea is that the "treatment" or new 2ndary schools should only affect the young cohort living in treatment villages. This generates the difference-in-difference design.

Key Variables:
treat = 1 if a new secondary school is built in your village
ycohort = 1 if person is aged 6 to 16 in 1985
ycohortxtreat = interaction term: ycohort X treat (use this for question 8)
primary = 1 if person completed primary school
electric = 1 if electricity in village ; **pipwater** = 1 if piped water in village ; **distcapital** = distance of village from capital
ocohort = 1 if person is aged 21-41 in 1985

1 The data collection effort was also sponsored by the following organizations: the Danish Agency for Development Assistance (DANIDA), the United States Agency for International Development (USAID), the University of Dar es Salaam, and the Economic Development Initiative (E.D.I., Tanzania).

* Indicates the question is optional and not is not required to get credit for the problem set.

Questions

Part I: Summary Statistics

Let's first look at summary statistics for the full sample. This will give you an idea of the characteristics of our population of interest. Then we will examine whether there are differences between treatment and control villages.

Q1) Present individual summary statistics for the study sample. What is the average age? What percentage of males are in the study? What is the average education?
Q2) Now present these same statistics for treatment villages vs. control villages. Are there any differences? (hint use: sum [variables] if treat== 0 or 1)
Q3*) 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 (using a p-value of .05 as the statistically significant threshold)? Are you concerned by any of the differences? How could they affect the analysis? (hint use: "ttest (variable), by(treat)")
Q4) Present community summary statistics for treatment villages vs. control villages. Are there any differences? Look at access to electricity, piped water, and distance from the capital.
Q5*) 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 (using a p-value of .05 as the statistically significant threshold)?
Q6) ANALYSIS: Based on your answer to either question 4 or 5, do you think that treatment villages are different from control villages? Why do you think secondary schools were built in treatment villages?

Part II: Simple Difference

Q7) Now we can simple compare young cohorts in treatment vs. control villages. Run the following regression only for the young cohort group. (Hint: reg [outcome] [variable] if ycohort==1)

$$Primary = \alpha + \beta(Treat) + \epsilon$$

What is your estimate of β ? Is it statistically significant (at the 5% level)?
Q8) ANALYSIS: Based on this result from question 7, can we say that constructing new 2ndary schools has a direct impact on primary school completion? Or can we attribute this to other factors (hint: use analysis from Q6).
Q9*) Do the regression from Q7 but add additional controls. Does your estimate of β change?
Q10*) Do the regression from Q7 but this time limit it only to the older cohort group. What is your estimate of β ? Is it statistically significant (at the 5% level)? How does it compare to your answer in Q7?

Part III: Diff-in-Diff comparing means

Q6) We are now just going to get the mean value for each cohort group.
- a) What percentage of young cohorts completed primary school in treatment villages?
- b) What percentage of old cohorts competed primary school in treatment villages?
- c) What percentage of young cohorts completed primary school in control villages?
- d) What percentage of old cohorts completed primary school in control villages?

Hint: use sum [variable] if ycohort==1 & treat== 0 or 1, or sum [variable] if ocohort==1 & treat== 0 or 1,

Q7) Now you can fill in the table below. Use your answers from Q6 (a), (b), (c), (d) and plug them into the table below. Take the differences from the treatment village vs. the control village, and you will get a naive diff-in-diff estimation of the effect of having a new 2ndary school built in your village.

* Indicates the question is optional and not is not required to get credit for the problem set.

	Treatment Village	Control Village	
Young Cohort	(a)	(c)	
Old Cohort	(b)	(d)	
First-Difference	(a)-(b)	(c)-(d)	Diff-in-Diff = [a-b]-[c-d]

Part IV: Diff-in-Diff in a regression framework

Q8) We will now estimate the treatment effect using a standard diff-in-diff regression.

$$Primary = \alpha + \beta 1(Treat) + \beta 2(Young\ Cohort) + \beta 3(Treat\ X\ Young\ Cohort) + \epsilon$$

What is your estimate of $\beta 3$? Is it statistically significant (at the 5% level)? How does it compare to your response in Q7?

Q9) ANALYSIS: Based on your analysis from Parts I, II, III, IV do you think building new secondary schools is effective at increasing primary school completion rates? What are some potential problems with the above analysis?

Part V*: Robustness (Optional Section)

You can only do this section with the large STATA data set. As a test of your research design above, you can look at two cohort groups that you know should not be affected by the treatment. Redefine the young cohort group as people aged (21-41) and the old cohort group as (42-61). Run the same regression from Q8.

HINT: I defined cohort2 = 1 to be the group that only includes people aged 21-61. Your regression will need to use this variable as a qualifer. For example, the regression will look like (reg [outcome] [variable] if cohort2==1]. I also defined ocohort2 = 1 if you are aged 42-62 in 1985.

What is your estimate of $\beta 3$? Is it statistically significant (at the 5% level)?

If you did find an effect, what would this tell you about the diff-in-diff research design? If you did not find an effect, does it give the diff-in-diff design more credibility?

* Indicates the question is optional and not is not required to get credit for the problem set.